SPATIOTEMPORAL GEOSTATISTICAL METHODS FOR EXPOSURE AND EPIDEMIOLOGICAL ANALYSES OF GROUNDWATER NITRATE AND RADON

Kyle Philip Messier

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Environmental Sciences and Engineering of the Gillings School of Global Public Health.

Chapel Hill
2015

Approved by:

Marc L. Serre

Rebecca C. Fry

Jacqueline A. MacDonald Gibson

Lawrence E. Band

Greg W. Characklis

ABSTRACT

Kyle P Messier: Spatiotemporal Geostatistical Methods for Exposure and Epidemiological Analyses of Groundwater Nitrate and Radon
(Under the Direction of Marc L. Serre)

Exposure assessment and dose-response characterization are critical steps in the risk assessment of an environmental contaminant with potential human health effects. There are many established methods to conduct exposure assessments and to characterize the dose-response relationship between a contaminant of concern and a health outcome; however, many require extensive time and monetary resources that are becoming increasingly limited. Geostatistical methods are attractive approaches due to their cost-effective implementation and clear physical interpretations. Land use regression (LUR) is a type of geostatistical method that uses spatially-based explanatory variables to model outcomes using classical regression methods. Bayesian Maximum Entropy (BME) is a geostatistical framework for incorporating measurements as well as various knowledge bases in a logical and theoretically sound manner to produce estimates for variables of interest at unmonitored locations. This work advances these spatiotemporal geostatistical methods in the following three studies: 1) An exposure assessment of groundwater nitrate (NO_3^-), a biological nutrient with natural and anthropogenic sources that in excess has deleterious effects on human and ecological health; 2) An exposure assessment of groundwater radon (^{222}Rn), a naturally occurring gas with radioactively discharged alpha particles that are known human carcinogens; and 3) An epidemiological analysis of the association between groundwater ^{222}Rn exposure and lung and stomach cancer incidence.

First, we develop a nonlinear LUR model and then integrate the model into the BME framework to produce the first space/time exposure estimates of groundwater NO_3^- concentrations across a large domain with a cross-validation r^2 of 0.74. Second, an exposure model for point-level groundwater ^{222}Rn is developed with anisotropic geological and uranium-

based explanatory variables resulting in a cross-validation $r^2$of 0.46. Lastly, we utilize the LUR-BME exposure model for ^{222}Rn to investigate associations with lung and stomach cancer at multiple spatial scales. It is the first epidemiological analysis of the association between groundwater ^{222}Rn exposure and lung cancer, moreover with a significant and positive association; and the first to find a positive association between groundwater ^{222}Rn and stomach cancer. This body of research provides advances in exposure assessment and dose-response methodology and practical real-world examples that can be used as resources for future cost-effective protection of public health.

For my late mother

Janice Irwin Messier

Loving, Beautiful, and Creative

ACKNOWLEDGEMENTS

I would first like to thank my advisor, Dr. Marc L. Serre, for bringing me into his lab and providing excellent mentorship through both a Master's and PhD. We had many a fruitful discussions ranging from technical statistics to abstract philosophy and international soccer matches. I am grateful to call Marc a mentor, a colleague, and a friend.

I would also like to thank my committee members, Dr. Rebecca C. Fry, for valuable research advice and for providing PhD funding for two years; Dr. Jacqueline A. MacDonald Gibson, for valuable input and research advice; Dr. Lawrence E. Band, for research advice and his excellent course covering the excitable research area of nitrogen; and Dr. Gregory W. Characklis, for reviewing and editing many papers and abstracts. I would also like to thank Dr. Leena Nylander-French for providing funding for two years during my PhD.

I would like to thank my friends in the department for making my time as a graduate student not only a valuable time professionally, but also socially. It was great to have many friends in the window-less, tornado bunker better known as the Rosenau basement. And of course I cannot forget our amazingly mediocre intramural basketball team. I still feel sorry for the referees.

Lastly, I would like to thank my family for all of their tremendous support. My dad, who has always been a proponent of academia and despite my better judgments, had some helpful advice along the way. To my step-mom Vicky, who is the definition of strength and perseverance. To my siblings and their families, Michael and Sarah, Mandy and Brian, and Matt, Courtney, Stella, and Holden: They are always supportive and I cannot think of better family to have. My newest family has been extremely supportive and engaging as well: Brent, Diane, David, and Jonathan Rager. And last, but not least, my biggest fan and the love of my life, my wife, Julia Rager... I mean Dr. Julia Rager.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

In order to protect the public from harmful contaminants in the environment, public health scientists conduct risk assessments, which have four basic steps: 1) risk identification, 2) exposure assessment, 3) dose-response characterization, and 4) health-risk characterization. There are many established methods to conduct exposure assessments and to characterize the dose-response relationship between a contaminant of concern and a health outcome; however, many require extensive time and monetary resources that are becoming increasingly limited. Developing geostatistical methods that utilize publicly available data to conduct risk assessment steps not only further develop our understanding of the contaminant of concern and protect public health, but also increase the returns on public resources spent on environmental and public health monitoring.

Understanding the risk of groundwater nitrate (NO_3^-) and radon (^{222}Rn) exposure is important because they are potential and known human carcinogens, respectively. The three studies in this work address the need for exposure assessment for groundwater NO_3^- and ^{222}Rn and the dose-response characterization for ^{222}Rn. The goals of this work are to further develop the spatiotemporal geostatistical methods that can utilize publicly available environmental and human health data, and to apply them to the novel assessment of groundwater NO_3^- and ^{222}Rn in North Carolina.

Nitrate Background

Nitrate (NO_3^-) is a biological nutrient with natural and anthropogenic sources that in excess has deleterious effects on human and ecological health[1]. Nitrate is part of the complex global nitrogen cycle (Figure 0.1), with natural sources including biological nitrogen fixation in grasslands and forests, bacterially mediated nitrification, and mineralization of organic nitrogen. Human derived or anthropogenic sources contribute at least twice as much to the presence of reactive nitrogen including NO_3^- in the environment compared to natural sources, which is largely due to agricultural development and the Haber-Bosch fertilizer synthesis process[2]. The broad categories of anthropogenic nitrate are agriculture fertilizer use, biological fixation of

cultivated crops, human and animal waste, combustion of fossil fuels including stationary and mobile sources, and other industrial processes.

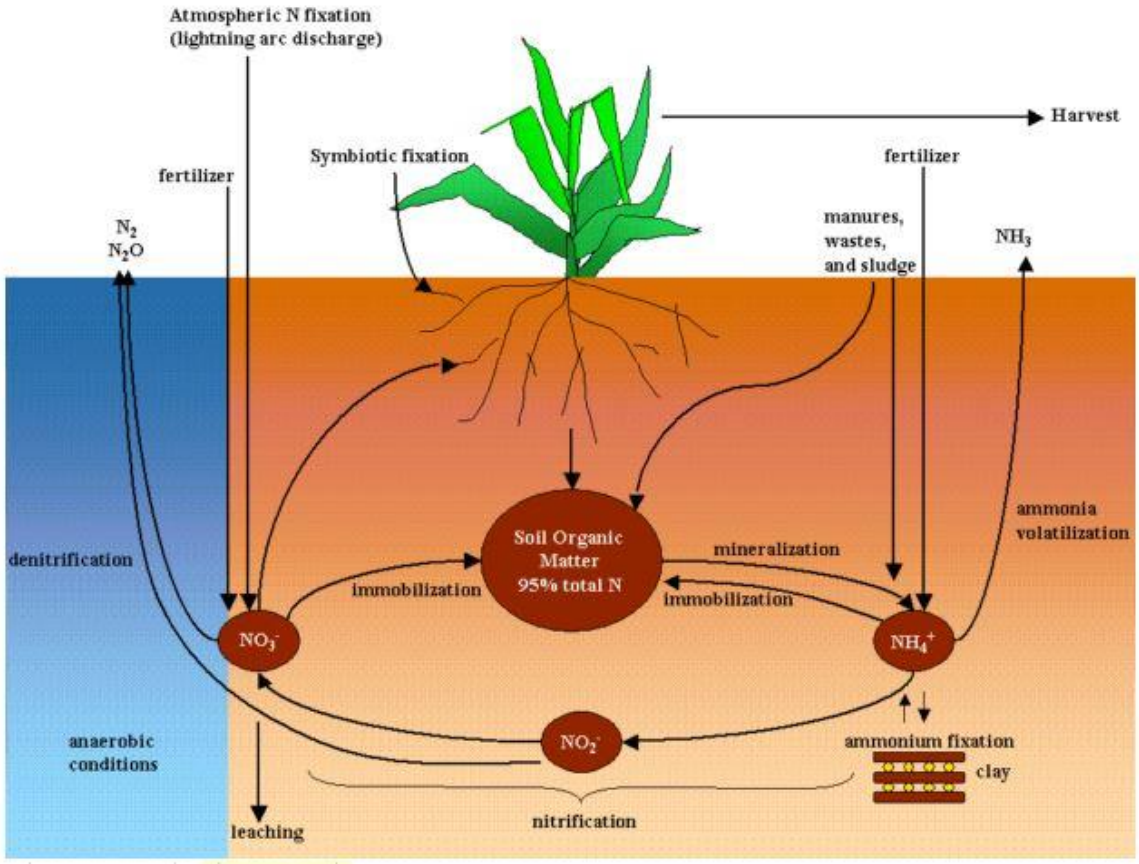


Figure 0.1. The nitrogen cycle. Figure verbatim from the Soil Water and Assessment Tool theoretical documentation, Neitsch et al. 2009 [3].

Exposure to NO_3^- can cause many deleterious health effects in humans. For instance, infants exposed to NO3- can develop methemoglobinemia, or blue baby syndrome[4]. This adverse endpoint is the basis of the 10 mg/L maximum contaminant level (MCL) in drinking water[5]. More recent studies have found associations between NO_3^- exposures at levels lower than the current MCL and cancers including colon[6], bladder[7], and Non-Hodgkin's Lymphoma[8]. Ecological effects resulting from excess NO_3^- in the environment include eutrophication of waterways, harmful algal blooms, and fish kills among others [9–11].

Radon Background

Radon (^{222}Rn) is a naturally occurring radioactive, inert, colorless, and odorless gas that is a daughter product of Uranium-238 and has a half-life of 3.83 days (Figure I2).

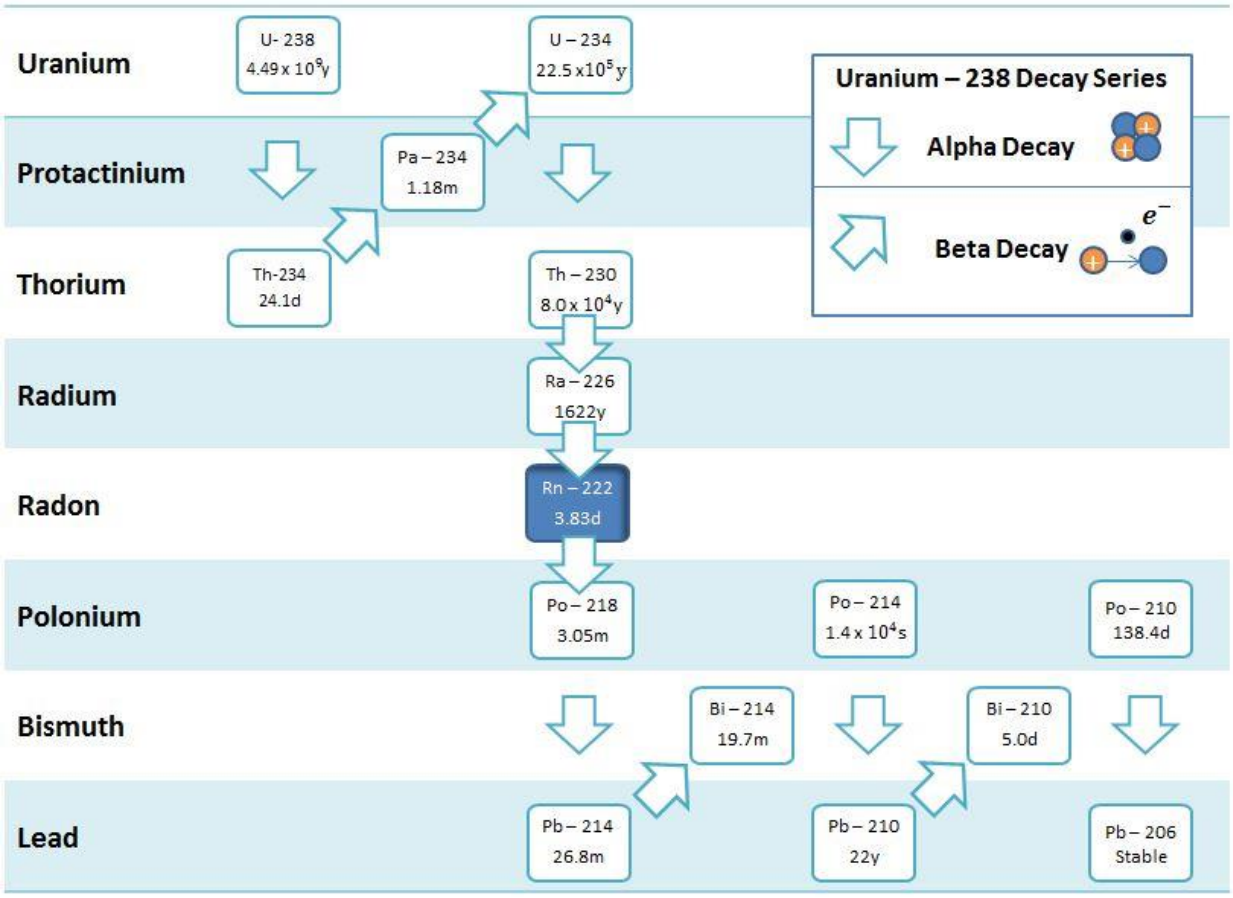


Figure 0.2. Uranium-238 decay series with Radon, the radionuclide of interest, highlighted in blue. The top row of each box is the element symbol and isotope number. The second row is the half-life time. Figure modified from Hall et al., 1985. Alpha decay is the release of 2 protons and two neutrons (a helium atom). Beta decay is the release of an electron and a proton changing to a neutron. y refers to years, d to days, m to minutes, and s to second.

^{222}Rn is found naturally in the soil, rocks, water, and air worldwide. ^{222}Rn and its daughter products or progeny produce ionizing radiation in the form of alpha and beta decay (Figure 0.2), which are known human carcinogens[12–14]. Outdoor air ^{222}Rn levels are generally very low; however, when ^{222}Rn enters a residential home, its concentration can increase to levels that may lead to adverse health effects [12]. Inhalation of indoor air contaminated with ^{222}Rn can lead to a significant increased risk of lung cancer morbidity in both never-smokers and smokers [14–16]. Exposure to ^{222}Rn is likely the second leading cause of lung cancer after smoking in the US[16–18]. Important routes of inhalation exposure result from ^{222}Rn gas directly escaping from soil and rock and accumulating in the indoor environment; however, ^{222}Rn can also degas from untreated

groundwater used for showering, dishwashing, and clothes washing resulting in exposures in direct vicinity to the breathing zone [19,20].

Land Use Regression

Land use regression (LUR) is a common statistical approach used for exposure assessments, which was introduced in the EU-funded SAVIAH (Small Area Variations in Air quality and Health) project in 1997[21]. Since LUR was introduced, over a hundred studies implemented LUR to assess exposure of air quality[22–24] and water quality contaminants[25–28]. LUR uses spatially based explanatory variables to model outcomes using classical regression methods. Examples of LUR explanatory variables include land use/land cover (LULC), altitude, river networks, road networks, and point source locations to name just a few. The major benefits of LUR include: 1) Simplicity as a regression-based approach; 2) The plethora of explanatory data sources with the increase of geographic information systems (GIS) and satellite databases; and 3) Physical interpretations of explanatory variables.

A LUR follows a standard regression format as follows:

$$Y_i = \beta_0 + \sum_{l=1}^{L} \beta_l X_i^l + \varepsilon_i \tag{0.1}$$

where Y_i is the outcome or dependent variable of interest at data point i, X_i^l is explanatory or independent variable l at point i, β_l is the regression coefficient for variable X^l, β_0 is the regression equation constant, ε_i is the error term for point i, and the summation represents the ability to include multiple explanatory variables. The LUR implementation of the regression equation is aided by including a spatial and/or time parameter dependency as follows:

$$Y_i(\boldsymbol{s}, t) = \beta_0 + \sum_{l=1}^{L} \beta_l X_i^l(\boldsymbol{s}, t) + \varepsilon_i \tag{0.2}$$

where $Y_i(\boldsymbol{s}, t)$ is the dependent variable for point i at spatial location $\boldsymbol{s}$ and temporal location t, and $X_i^l(\boldsymbol{s}, t)$ is explanatory variable l at point i at the same spatial location $\boldsymbol{s}$ and temporal location t. Model coefficients are determined with same techniques available for ordinary linear regression such as ordinary least squares (OLS) and generalized least squares (GLS). In multivariable models, the final model may be selected with traditional statistical techniques such as forward selection, backwards selection, step-wise selection, lasso, and least angle regression;

or with techniques developed specifically for LUR such A Distance Decay Regression Selection Strategy[22].

Bayesian Maximum Entropy

Bayesian Maximum Entropy (BME) is a modern spatiotemporal geostatistical framework for incorporating measurements as well as various knowledge bases in a logical and theoretically sound manner to produce estimates of variables of interest at unmonitored locations[29]. BME therefore is an extremely valuable tool that can be used to produce information of chemical levels and disease rates through space and time by making efficient use of available monitoring resources. BME consists of three epistemological stages known as the *prior*, *meta-prior*, and *posterior* stages.

It is important to first define the notation and some basic concepts for discussing these BME stages: The notation for a single random variable Z in capital letter, its realization, z, in lower case; and vectors and matrices in bold faces, e.g. $\boldsymbol{Z} = [Z_1, \dots, Z_n]^T$ and $\boldsymbol{z} = [z_1, \dots, z_n]^T$. Let $\chi(\boldsymbol{p})$ be the space/time random field (S/TRF) describing the distribution of a variable of interest across space and time, where $\boldsymbol{p} = (\boldsymbol{s}, t)$, $\boldsymbol{s}$ is the space coordinate and t is time.
The *prior* stage (Figure I3, Panel A) consists of gathering information or knowledge about the space/time distribution of the variable of interest and compiling it into the general knowledge base,$G - KB$. In the *prior* stage the $G - KB$ is mathematically a set of integral equations, which represent constraints on the space/time distribution of $\boldsymbol{\chi}_{map}$:

$$E[\mathcal{g}_\alpha] = \int f_\chi(\boldsymbol{\chi}_{map}) \boldsymbol{\mathcal{g}_\alpha}(\boldsymbol{\chi}_{map}) d\boldsymbol{\chi}_{map} \tag{0.3}$$

where $\alpha = 0,1, \dots, N_c$ is the number of suitable constraint functions, $f_\chi(\boldsymbol{\chi}_{map})$ is the probability distribution function (PDF) of $\boldsymbol{\chi}_{map} = [x_1, \dots, x_m x_k]^T$, $\mathcal{g}_\alpha(\boldsymbol{\chi}_{map})$ are known functions of $\boldsymbol{\chi}$, and $E[.]$ is the expected value operator**.** Functions that describe the space/time distribution include the mean trend, covariance, higher-order moments such as the trivariance, regression equations, and stochastically represented mechanistic equations (i.e. physical laws). The prior PDF describing the space/time distribution of $\boldsymbol{\chi}_{map}$ is given by:

$$f_\chi(\boldsymbol{\chi}_{map}) = \exp\left(\mu_0 + \sum_{\alpha=1}^{N_c} \mu_\alpha \mathcal{g}_\alpha(\chi_{map}) \right) \tag{0.4}$$

where $\mu_\alpha, (\alpha = 0,1, \dots, N_c)$ are Lagrange coefficients that are solved for via the system of $N_c + 1$ equations, with the first equation, $\int f_\chi(\boldsymbol{\chi}_{map})\, d\boldsymbol{\chi}_{map} = 1$, as the normalization constraint.

In the *meta-prior* stage one gathers and organizes all of the information that can be explicitly incorporated into BME in the site-specific knowledge base, $S - KB$. This entails identifying hard data, S: $\boldsymbol{\chi}_{hard}$, or data without measurement error (Figure 0.3, Panel B); soft data, S: $\boldsymbol{\chi}_{soft}$, or data with measurement error that is expressed mathematically as a distribution function. Soft data can be data that has inherent error quantified in its measurements (Figure 0.3, Panel C) or data that is the result of a model prediction with confidence bounds (Figure 0.3, Panel D). An important part of BME is that the soft data can be represented with any distributional form, or is not limited to linear, normal/Gaussian distributions. For example, possible soft data distributions in addition to Gaussian are interval, truncated-Gaussian, or cumulative distribution function. This flexibility in distributions allow for more accurate modeling of environmental contaminants. In practice, the information in the *meta-prior* stage is often used in the *prior* stage to empirically derive the functions for the general knowledge base such as the mean trend and covariance.

The *posterior* stage (Figure 0.3, Panel E) updates the prior PDF, $f_\chi(\boldsymbol{\chi}_{map})$, with site-specific knowledge from the *meta-prior* stage using Bayesian conditionalization, which is essentially the marginal PDF of $f_\chi(\boldsymbol{\chi}_{map})$ with respect to $\boldsymbol{\chi}_{soft}$ or $\boldsymbol{\chi}_{hard}$, depending on the scenario. Given the scenario of hard data and probabilistic soft data, the BME posterior PDF for a given estimation point x_k is given by:

$$f_k(\boldsymbol{\chi_k}) = A^{-1} \int_D f_\chi(\boldsymbol{x_{hard}}, \boldsymbol{x_{soft}}, x_k) f_\chi(\boldsymbol{x_{soft}}) d\boldsymbol{\chi}_{soft} \tag{0.5}$$

where

$$A = \int_D d\chi_k \int_D f_G(\boldsymbol{x_{soft}}, \boldsymbol{x_{hard}}, x_k) d\boldsymbol{\chi}_{soft} \tag{0.6}$$

is the normalization constraint, and $\int_D(.)$ is the integral over the domain of the soft data.

When the general knowledge base consists of the mean trend and covariance only and the site-specific data contains only hard data or soft data with a Gaussian measurement error, then BME reduces to the Kriging estimator. As BME currently stands in its numerical implementation, secondary information above the space/time distribution of $\boldsymbol{\chi}$ is implemented

through the mean trend as opposed to additional g_α functions in the general knowledge base at the *prior* stage.

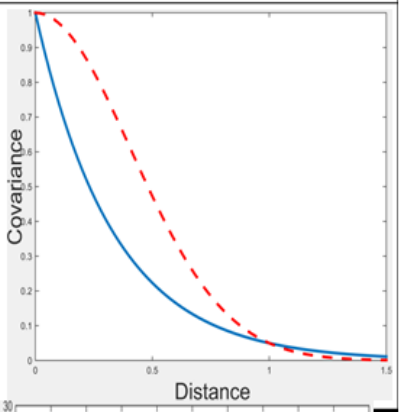
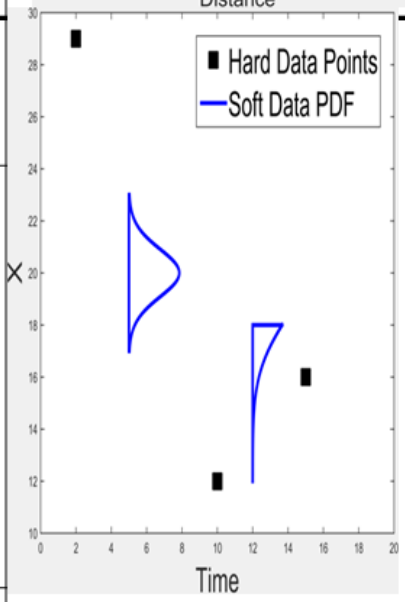
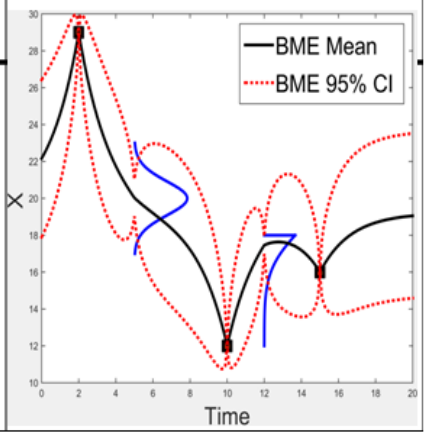
Stage	Step		Formula	Illustration
Prior Stage	A	Prior PDF derived from general knowledge base, G. Functions include the mean trend, covariance, higher-order spatial moments, or stochastic physical laws. Relationships are determined analytically or empirically with existing dataset.	Space/Time Random Field (S/TRF) $X(p) = X(s,t)$ is a random variable that is a function of location $s=(s_1,s_2)$ and time t. Mean Trend: $m_x(\boldsymbol{p}) = E[X(\boldsymbol{p})]$ space/time covariance : $C_x(\boldsymbol{p},\boldsymbol{p}')$ $= E\left[\left(\chi(\boldsymbol{p}) - m_\chi(\boldsymbol{p})\right)\left(\chi(\boldsymbol{p}') - m_\chi(\boldsymbol{p}')\right)\right]$	
Meta-Prior Stage	B	Hard data (if any) identified in the site-specific knowledge base, S.	S: $\chi_{hard}, \sigma^2_{hard} = 0$	
	C	Soft data, or data with measurement error expressed as a probability distribution, is identified in the site-specific knowledge base, S. Soft data can also be obtained from model predictions, such as land use regression.	S: $\chi_{soft} = f_{x_{soft}}(.)$ possible soft data distributions are interval, Gaussian, truncated-Gaussian, cumulative distribution function, or any other probabilistic form.	
	D	Hard and Soft data may be together	S: $\chi_{data} = (\chi_{hard}, \chi_{soft})$	
Posterior Stage	E	In light of site-specific knowledge base, the general knowledge base is updated by means of *Bayesian Conditionalization* leading to the BME PDF for the map. Uncertainty is zero at locations with hard data and reduced at soft data.	$f_k(\chi_k)$ $= A^{-1}\int_D f_G(\boldsymbol{x_{hard}}, \boldsymbol{x_{soft}}, x_k) \ldots$ $f_S(\boldsymbol{x_{soft}}) d\boldsymbol{\chi_{soft}}$ $where\ A = \int_D d\chi_k \ldots$ $\int_D f_G(\boldsymbol{x_{soft}}, \boldsymbol{x_{hard}}, x_k) d\boldsymbol{\chi_{soft}}$ $\int_D(.)$ is the domain of the soft data.	

Figure 0.3. An illustration of BME methodology (Revised from LoBuglio et al. 2007[30]).

Project Themes

The goal of this work is to advance the spatiotemporal geostatistical methods that are utilized in exposure assessments and epidemiological studies; and to demonstrate methodological improvements in the modeling of groundwater NO_3^- and ^{222}Rn in North Carolina. Methodological themes that are present in this work include: 1) Land Use Regression model development for groundwater contaminants; 2) Implementation of space/time BME estimation for groundwater contaminants; 3) Integration of Land Use Regression models into the Bayesian Maximum Entropy framework; 4) Model selection procedures in large variable space problems; 5) Quantitative comparisons of model results arising from differences in spatial scales of independent and dependent variables; and 6) Quantitative comparisons between the current state of science in exposure estimates and dose-response characterization and the novel developments from this work.

Accurate and precise exposure assessments are important for both nitrate and radon due to their significant human and ecological health risks. Not only is a detailed mapping of groundwater nitrate important from a biogeochemical perspective[31], but it also poses known and potential human health effects, including cancers[6,7], that need quality exposure assessments for further study. Similarly, radon has known and potential human carcinogenetic effects. This dissertation addresses that need by providing a framework for more accurate exposure assessment, which is shown with two case studies of nitrate and radon. Additionally, it will be shown that the exposure assessments can then be utilized in an epidemiological analysis to help elucidate the association between exposures and a response, which in this case is radon and cancers of the stomach and lung. In short, this work comprises two exposure assessments and one dose-response characterization through an epidemiological analysis.

Dissertation Organization

This dissertation is organized into three chapters, with each chapter formatted as a publishable quality manuscript. First, a state-wide exposure assessment of the deleterious human and environmental contaminant of groundwater nitrate is conducted. A nonlinear LUR and geostatistical method is implemented, which incorporates information on nitrate sources, and attenuation and transport factors. This manuscript was accepted into the journal *Environmental Science and Technology* in August of 2014. Second, similar to the nitrate exposure assessment, a linear LUR model is used to estimate groundwater radon state-wide; however, the LUR utilizes

information pertinent to radon including lithological and uranium data. This manuscript has been submitted to the journal *Water*. Third, the exposure assessment of groundwater radon is used in the dose-response characterization of groundwater radon to the health outcomes of stomach and lung cancer via an epidemiological analysis at the ecological and the address-level scales. We plan on submitting this manuscript to the journal *International Journal of Epidemiology* in the near future.

Each chapter in this dissertation, including this introduction, has independent reference sections. Additionally there is an overall dissertation conclusion that summarizes all three chapter results, discusses the public health relevance of the overall work, and projects future research potential.

REFERENCES

(1) US Environmental Protection Agency. Basic Information About Nitrate in Drinking Water http://water.epa.gov/drink/contaminants/basicinformation/nitrate.cfm (accessed Nov 1, 2012).

(2) Doering, O. C. I.; Galloway, J. N.; Theis, T. L.; Aneja, V.; Boyer, E.; Cassman, K. G.; Cowling, E. B.; Dickerson, R. R.; Herz, W.; Hey, D. L.; et al. *Reactive Nitrogen in the United States: An Analysis of Inputs, Flows, Consequences, and Management Options. EPA-SAB-11-013*; Washington D.C., 2011.

(3) Neitsch, S. L.; Arnold, J. G.; Kiniry, J. R.; Williams, J. R. Soil & Water Assessment Tool Theoretical Documentation. **2009**.

(4) Comly, H. H. Cyanosis in infants caused by nitrates in well water. *J. Am. Med. Assoc.* **1945**, *129*, 112–116.

(5) Spalding, R. F.; Exner, M. E. Occurrence of Nitrate in Groundwater—A Review. *J. Environ. Qual.* **1993**, *22*, 392–402.

(6) De Roos, A. J.; Ward, M. H.; Lynch, C. F.; Cantor, K. P. Nitrate in Public Water Supplies and the Risk of Colon and Rectum Cancers. *Epidemiology* **2003**, *14*, 640–649.

(7) Weyer, P. J.; Cerhan, J. R.; Kross, B. C.; Hallberg, G. R.; Kantamneni, J.; Breuer, G.; Jones, M. P.; Zheng, W.; Lynch, C. F. Municipal Drinking Water Nitrate Level and Cancer Risk in Older Women : The Iowa Women's Health Study. *Epidemiology* **2001**, *12*, 327–338.

(8) Ward, M. H.; Mark, S. D.; Cantor, K. P.; Weisenburger, D. D.; Correa-Villasenor, A.; Zahm, S. H. Drinking Water Nitrate and the Risk of Non-Hodgkin ' s Lymphoma. *Epidemiology* **1996**, *7*, 465–471.

(9) Paerl, H. W. Coastal eutrophication and harmful algal blooms : Importance of atmospheric deposition and groundwater as " new " nitrogen and other nutrient sources. *Limnol. Oceanogr.* **1997**, *42*, 1154–1165.

(10) Zhou, M.; Shen, Z.; Yu, R. Responses of a coastal phytoplankton community to increased nutrient input from the Changjiang (Yangtze) River. *Cont. Shelf Res.* **2008**, *28*, 1483–1489.

(11) Smith, V. H.; Tilman, G. D.; Nekola, J. C. Eutrophication: impacts of excess nutrient inputs on freshwater, marine, and terrestrial ecosystems. *Environ. Pollut.* **1999**, *100*, 179–196.

(12) WHO. *WHO guidelines for drinking-water quality.*; 2011.

(13) Krewski, D.; Lubin, J. H.; Zielinski, J. M.; Alavanja, M.; Catalan, V. S.; Field, R. W.; Klotz, J. B.; Letourneau, E. G.; Lynch, C. F.; Lyon, J. I.; et al. Residential Radon and Risk of Lung Cancer. *Epidemiology* **2005**, *16*, 137–145.

(14) Lubin, J. H.; Boice, J. D. Lung cancer risk from residential radon: meta-analysis of eight epidemiologic studies. *J. Natl. Cancer Inst.* **1997**, *89*, 49–57.

(15) Field, R. W.; Smith, B.; Steck, D.; Lynch, C. F. Residential radon exposure and lung cancer: variation in risk estimates using alternative exposure scenarios. *J. Expo. Anal. Environ. Epidemiol.* **2002**, *12*, 197–203.

(16) Kendall, G. M.; Smith, T. J. Doses to organs and tissues from radon and its decay products. *J. Radiol. Prot.* **2002**, *22*, 389–406.

(17) Campbell, T.; Mort, S.; Fong, F.; Crawford-Brown, D.; Vengosh, A.; Cornell, E.; Field, W. R. *North Carolina Radon-in-Water Advisory Committee Report*; Raleigh, North Carolina, 2011.

(18) National Research Council. *Risk Assessment of Radon in Drinking Water*; Washington D.C., 1999.

(19) Vinson, D. S.; Campbell, T. R.; Vengosh, A. Radon transfer from groundwater used in showers to indoor air. *Appl. Geochemistry* **2008**, *23*, 2676–2685.

(20) Fitzgerald, B.; Hopke, P. K. Experimental Assessment of the Short- and Long-Term Effects of Rn from Domestic Shower Water on the Dose Burden Incurred in Normally Occupied Homes. **1997**, *31*, 1822–1829.

(21) Briggs, D. J.; Collins, S.; Elliott, P.; Fischer, P.; Kingham, S.; Lebret, E.; Pryl, K.; Van Reeuwijk, H.; Smallbone, K.; Van Der Veen, A. Mapping urban air pollution using GIS: a regression-based approach. *Int. J. Geogr. Inf. Sci.* **1997**, *11*, 699–718.

(22) Su, J. G.; Jerrett, M.; Beckerman, B. A distance-decay variable selection strategy for land use regression modeling of ambient air pollution exposures. *Sci. Total Environ.* **2009**, *407*, 3890–3898.

(23) Raaschou-Nielsen, O.; Andersen, Z. J.; Beelen, R.; Samoli, E.; Stafoggia, M.; Weinmayr, G.; Hoffmann, B.; Fischer, P.; Nieuwenhuijsen, M. J.; Brunekreef, B.; et al. Air pollution and lung cancer incidence in 17 European cohorts: prospective analyses from the European Study of Cohorts for Air Pollution Effects (ESCAPE). *Lancet. Oncol.* **2013**, *14*, 813–822.

(24) Reyes, J. M.; Serre, M. L. An LUR/BME Framework to Estimate PM2.5 Explained by on Road Mobile and Stationary Sources. *Environ. Sci. Technol.* **2014**, *48*, 1736–1744.

(25) Hoos, A. B.; Mcmahon, G. Spatial analysis of instream nitrogen loads and factors controlling nitrogen delivery to streams in the southeastern United States using spatially referenced regression on watershed attributes (SPARROW) and regional classification frameworks. *Hydrol. Process.* **2009**, *23*, 2275–2294.

(26) McLay, C. D.; Dragten, R.; Sparling, G.; Selvarajah, N. Predicting groundwater nitrate concentrations in a region of mixed agricultural land use: a comparison of three approaches. *Environ. Pollut.* **2001**, *115*, 191–204.

(27) Aelion, C. M.; Conte, B. C. Susceptibility of residential wells to VOC and nitrate contamination. *Environ. Sci. Technol.* **2004**, *38*, 1648–1653.

(28) Messier, K. P.; Akita, Y.; Serre, M. L. Integrating address geocoding, land use regression, and spatiotemporal geostatistical estimation for groundwater tetrachloroethylene. *Environ. Sci. Technol.* **2012**, *46*, 2772–2780.

(29) Christakos, G.; Li, X. Bayesian Maximum Entropy Analysis and Mapping : A Farewell to Kriging Estimators ? *Math. Geol.* **1998**, *30*, 435–462.

(30) LoBuglio, J. N.; Characklis, G. W.; Serre, M. L. Cost-effective water quality assessment through the integration of monitoring data and modeling results. *Water Resour. Res.* **2007**, *43*, 1–16.

(31) Rivett, M. O.; Buss, S. R.; Morgan, P.; Smith, J. W. N.; Bemment, C. D. Nitrate attenuation in groundwater: a review of biogeochemical controlling processes. *Water Res.* **2008**, *42*, 4215–4232.

CHAPTER 1

Nitrate Variability in Groundwater of North Carolina using Monitoring and Private Well Data Models

Kyle P. Messier[†], *Evan Kane*[‡], *Rick Bolich*[‡], *Marc L. Serre*[†]*

Authors' Affiliation:

† Department of Environmental Science and Engineering, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC

‡ North Carolina Department of Environment and Natural Resources, Division of Water Resources

***Corresponding Author:**

Marc L. Serre

Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, University of North Carolina, 1303 Michael Hooker Research Center, Chapel Hill, NC 27599

Phone: (919) 966-7014 Fax: (919) 966-7911

Abstract

Nitrate (NO_3^-) is a widespread contaminant of groundwater and surface water across the United States that has deleterious effects to human and ecological health. This study develops a model for predicting point-level groundwater NO_3^- at a state scale for monitoring wells and private wells of North Carolina. A land use regression (LUR) model selection procedure is developed for determining nonlinear model explanatory variables when they are known to be correlated. Bayesian Maximum Entropy (BME) is used to integrate the LUR model to create a LUR-BME model of spatial/temporal varying groundwater NO_3^- concentrations. LUR-BME results in a leave-one-out cross-validation r^2 of 0.74 and 0.33 for monitoring and private wells, effectively predicting within spatial covariance ranges. Results show significant differences in the spatial distribution of groundwater NO_3^- contamination in monitoring versus private wells; high NO_3^- concentrations in the southeastern plains of North Carolina; and wastewater treatment residuals and swine confined animal feeding operations as local sources of NO_3^- in monitoring wells. Results are of interest to agencies that regulate drinking water sources or monitor health outcomes from ingestion of drinking water. Lastly, LUR-BME model estimates can be integrated into surface water models for more accurate management of non-point sources of nitrogen.

Introduction

Nitrate (NO_3^-) is a widespread contaminant of groundwater and surface water across the United States that has deleterious effects to human and ecological health[1,2]. The maximum contaminant level of 10 mg/L established by the U.S. Environmental Protection Agency was based on the prevention of methemoglobinemia in infants[3]; moreover, there is concern of many cancer types [4–6] and from lower concentration exposures[7]. Excessive NO_3^- inputs into the environment can result in adverse changes to ecosystems such as eutrophication and harmful algal blooms[8–10].

Protection of drinking water sources is mandated by the Safe Drinking Water Act; however, private well drinking water is unregulated in contrast to regulated public water systems[11]. In North Carolina where more than ¼ of the population relies on private wells for drinking water[12], quantifying potential exposures is important to protect public health. Monitoring programs such as the US Geological Survey's (USGS) National Water Quality Assessment (NAWQA) Program[13] and the NC Division of Water Resources (NC DWR) ambient monitoring program[14] are effective because they use consistent sampling and analytical methods, yet this water quality monitoring data is spatially and temporally sparse.

Land use regression[15–21](LUR) is a proven method that complements monitoring programs and provides effective means for water quality exposure assessments. Previous studies have related land use characteristics to NO_3^- contamination in surface waters [22–25] and groundwater. Additionally, regression-based methods have been implemented for estimating loading to surface waters[21,23,24]. In North Carolina, groundwater discharge to streams (baseflow) accounts for roughly two-thirds of annual streamflow in the Coastal Plains region of North Carolina[26] and may be contributing excess nutrient loads in streams[27]; however, current surface water models do not directly account for this large source of NO_3^- from baseflow.

For linear regression models, traditional statistical methods to select predictor variables include forward, backwards, and stepwise selection. These methods can lead to erroneous models with high multicollinearity when the candidate variables are related. However, for LUR model studies, model selection methods have been modified to accommodate the potential high multicollinearity from selection variables that differ only by a hyperparameter[16,19]. Additionally, lasso[28] and elastic net[29] regression are potential methods for selecting linear LUR models, but to the authors' knowledge has not been employed for LUR model selection. For nonlinear

regression, methods for model selection based on a large candidate variable space include stepwise logistic regression[30,31] and regression tree analysis which approximates nonlinear relationships [32,33]; still for continuous variable outcomes with nonlinear models, less rigorous methods for model selection have been developed. The number of candidate variables is generally consolidated to a tractable number through expert knowledge or single variable regression, and then various combinations of models are tested until one finds the best model in terms of a validation statistic like R^2 or Akaike Information Criterion (AIC)[15,21,24].

The advanced geostatistical method of Bayesian Maximum Entropy (BME) has also been shown to successfully estimate groundwater quality contaminants[19,34]. An advantage of BME is its ability to quantify spatial and temporal variability which is then used in the estimation process at unmonitored locations. BME, like all geostatistical methods, is data driven and can only provide reliable estimates within the vicinity of measured values. However, BME utilizes Bayesian epistemic knowledge blending to combine multiple sources of data, which has been successfully demonstrated with incorporation of deterministic mean trend functions into BME for groundwater[19].

Local spatial and temporal variability have lead previous studies to reduce NO_3^- variability with a combination of spatial smoothing and temporal averaging[15,35,36]. For instance, Nolan and Hitt spatially smoothed NO_3^- by taking watershed averages over their study time period, based on watersheds with an average size of approximately 2000 square-kilometers. They not only helped elucidate trends and potential explanatory variables, but they were able to explain a large percentage in the variability of spatially-smoothed NO_3^- with a r^2 of 0.80 for shallow aquifer NO_3^- and 0.77 for deep aquifer NO_3^-. However, this advantage of reducing groundwater NO_3^- variance is also a limitation because factors affecting spatially-smoothed and temporally averaged NO_3^- might not affect point-level NO_3^-, and vice-versa. Furthermore, since groundwater NO_3^- contains significant local variability, the need to provide local estimates of its variability naturally follows. Models developed for predicting spatially-smoothed and temporally averaged NO_3^- will likely not be successful in predicting observed, point-level NO_3^-.

The objectives of this study are to: 1) Develop a novel nonlinear regression model for spatial point-level and time-averaged groundwater NO_3^- concentrations in monitoring and private wells of North Carolina, 2) Produce the first space/time estimates of groundwater NO_3^- concentrations across a large study domain by integrating LUR models into the BME framework,

and 3) Compare space/time NO_3^- concentration models to the current standard of spatially averaged NO_3^- concentration models Two nonlinear models, whose form is adopted from Nolan and Hitt[15] with components that represent NO_3^-sources, attenuation, and transport, are created and selected with a new model selection framework for nonlinear regression models with correlated explanatory variables. We then integrate the LUR models into the BME framework to model space/time point-level NO_3^-. Results are of interest to agencies that regulate drinking water sources or that monitor health outcomes from ingestion of drinking water. Additionally, the results can provide guidance on factors affecting the point-level variability of groundwater $NO_3^-$and new resources for more accurate management of NO_3^- loads.

Methods

Nitrate Data

NO_3^-data across North Carolina are obtained from three data sources (Figure S1.1), which are detailed as follows:

North Carolina Division of Water Resources (NC-DWR) collects data near select permitted, dedicated Wastewater Treatment Residual (WTR) application fields via monitoring wells. The second source is USGS data obtained through the National Water Information System (NWIS). Well depth information is not linked directly to each monitoring well, although a subset of well depth information is available. Based on the subset with depth information, they have a mean depth of 33 feet with a standard deviation of 32 feet. Together, the NCDWR and USGS data represent shallow aquifer monitoring wells (n= 12,322), which hereafter will be referred to as "Monitoring Well data."

The last dataset of groundwater NO_3^-comes from private well data collected by the North Carolina Department of Health and Human Services (NC-DHHS). Groundwater $NO_3^-$was obtained and address geocoded using the same process outlined in Messier et al.[19]. Well depth information is not linked to water quality measurements, but a separate database on private well construction contains well depths. The mean depth is 95 feet with a standard deviation of 109 ft. This data will hereafter be referred to as "Private Well data" and this data is assumed to represent a deeper aquifer model of groundwater NO_3^- (n=22,067).

The median NO_3^- concentrations for the NC-DWR, USGS, and private well data are 1.30, 0.10, and 0.62 mg/L respectively. The means are 4.61, 6.14, and 1.66 mg/L respectively. The

percent observed above the detection limit is 79.7, 61.4, and 30.6 respectively. Additional basic statistics for the dataset are available in the supporting information (Table S1.1).

Spatial and Temporal Observation Scales

In this work we develop models for NO_3^- at three observation scales. The finer scale corresponds to the space/time *point-level* NO_3^- data, i.e. NO_3^- data as it is sampled. An intermediate observation scale corresponds to the *time-averaged* data, whereby NO_3^- at each well is averaged. The time-averaged data provides point-level spatial resolution, but no time variability. Finally, the coarser resolution observation scale corresponds to the *spatially-smoothed/time-averaged* data, which was obtained by spatially smoothing the time-averaged data using a 25 km exponential kernel function. We choose 25 km as it is approximately the average size of watersheds in many NAWQA groundwater studies[15,37]. While previous works over large study domains have developed models for spatially-smoothed/time average NO_3^- data, very few models, if any, have been developed for point-level NO_3^- data over large study domains. Our work therefore fills that knowledge gap.

Maximum Likelihood Estimation of Nitrate Distributions

Our notation for variables denotes a single random variable *Z* in capital letter, its realization, *z,* in lower case; and vectors and matrices in bold faces, e.g. $\boldsymbol{Z} = [Z_1, \dots, Z_n]^T$ and $\boldsymbol{z} = [z_1, \dots, z_n]^T$.

Due to the high percentage of nondetect (left-censored) data in both the monitoring well and private well databases, a maximum likelihood estimation (MLE) is used for the estimation of monitoring well and private well distribution parameters[38], which is assumed to follow a lognormal distribution. MLE can directly account for the nondetect values by modifying the likelihood equation, with the censored observations given by the cumulative distribution function (CDF) evaluated at the detection limit. The MLE equation then becomes[38]:

$$\mathcal{L}(\boldsymbol{z}|\mu,\sigma) = \left\{ \prod_{z_i | z_i \geq t_i} f_{\mu,\sigma}(z_i) \right\} * \left\{ \prod_{z_i | z_i \leq t_i} F_{\mu,\sigma}(t_i) \right\} \tag{1.7}$$

where $f_{\mu,\sigma}(z_i)$ denotes the normal probability distribution function (PDF) of log-transformed (natural log) point-level NO_3^-, z_i, with mean and standard deviation parameters μ and σ, and $F_{\mu,\sigma}(t_i)$ denotes the CDF of the distribution taken at the log of the detection limit t_i. The estimated distributions are used to quantify the extent of contamination in monitoring and private

wells and to handle nondetect data. For the regression analysis, the log- NO_3^- concentration of a measurement below detection limit t_i is assigned a value equal to the mean of the normal distribution N(μ,σ) truncated above log(t_i), whereas the geostatistical analysis can handle the full truncated normal distribution[19].

Spatial Explanatory Variables

Spatial explanatory variables representing possible groundwater NO_3^- sources, attenuation, and transport factors were constructed prior to model development. Potential variables are summarized below with details available in the supporting information (Table S1.2).

All of the explanatory variables have an inherent spatial distance parameter such as circular buffer radius or exponential decay range, which hereinafter is referred to as the *hyperparameter*. Each variable is calculated with multiple hyperparameter values since optimal distance is unknown a priori. In the final model selection process, a maximum of one hyperparameter value is allowed to be selected from each variable to avoid multicollinearity and effectively optimize the hyperparameter. The following variables adopted from Nolan and Hitt[15] are NO_3^- sources calculated as $\mathrm{Kg-NO_3^-/yr/ha}$ within a circular buffer: Sources include farm fertilizer, non-farm fertilizer, manure, and NO_3^- atmospheric deposition. Each National Landcover Database (NLCD) category is calculated as a percent within a circular buffer. On-site wastewater treatment plant variables, septic density and average nitrate loading, are created following the methods of Pradhan et al.[39] The following point sources are calculated as the sum of exponentially decaying contribution[19]: Wastewater treatment residual field application sites (WTR), swine confined animal feeding operations (CAFOs), poultry CAFOs, cattle farms, and wastewater treatment plants (WWTP). Mean slope in degrees and topographic wetness index[40] (TWI) are calculated within circular buffers. Water withdrawals in cubic meters per second are calculated using USGS water use estimates[12]. Lastly, population density is calculated within circular buffers from US Census block data assuming an even distribution of population per census block.

Nonlinear Regression Model Selection

In order to develop a LUR model for NO_3^- we adopt a similar nonlinear multivariable model implemented by Groundwater Vulnerability Assessment(GWAVA)[15] which is also similar to the surface water counterpart Spatially Referenced Regression On Watershed Attributes

(SPARROW) [21,23,24]. We partition explanatory variables into source, attenuation, and transport terms. Following Nolan and Hitt[15], the nonlinear multivariable model is constructed as follows:

$$z_i = \beta_0 + \left\{\sum_{k=1}^{K} \beta_k Y_i^{(k)}(\lambda_k)\right\} \exp\left\{\sum_{l=1}^{L} -\gamma_l Y_i^{(l)}(\lambda_l)\right\} \exp\left\{\sum_{m=1}^{M} \delta_m Y_i^{(m)}(\lambda_m)\right\} + \varepsilon_i \quad (1.8)$$

where z_i is the log–transform of NO_3^-concentration at point *i*, β_0 is the intercept, $Y_i^{(k)}(\lambda_k)$ is the *k*-th source predictor variable at point *i* with hyperparameter value λ_k, β_k is its source regression coefficient, $Y_i^{(l)}(\lambda_l)$ is the *l*-th attenuation predictor variable at point *i* with hyperparameter value λ_l, $\gamma_l$is its attenuation regression coefficient, $Y_i^{(m)}(\lambda_m)$ is the *m*-th transport predictor variable with hyperparameter value λ_m, δ_m is its transport regression coefficient, and ε_i is an error term. The model contains an additive, linear submodel for sources, and multiplicative exponential terms for the attenuation and transport variables that act directly on the source terms [15]. For example $Y_i^{(k)}(\lambda_k)$ may be equal to a land cover variable or a point source variable. The attenuation variables,$Y_i^{(l)}$, physically represent areas that are associated with removing $NO_3^-$from groundwater such as wetlands and histosol soil. The transport variables, $Y_i^{(m)}(\lambda_m)$, may be equal to any variable that effects the movement of NO_3^-in the groundwater such as the soil permeability and average slope. The attenuation variable coefficients,γ_l, are constrained to be negative allowing them to only decrease NO_3^- concentrations, while the transport variable coefficients,δ_m, are unconstrained allowing variables to increase or decrease NO_3^-concentrations.

We developed a nonlinear model regression model selection technique that accommodates variables that differ only by a hyperparameter and can be adapted for various nonlinear model forms. Our model selection procedure is essentially a nonlinear extension of A Distance Decay REgression Selection Strategy (ADDRESS)[16], since to the authors' knowledge there is not a regression selection strategy for nonlinear LUR. We implement Constrained Forward Nonlinear Regression with Hyperparameter Optimization (CFN-RHO) whose simple algorithm is as follows (Figure S1.2):

1) *Initialization:* Linear regression on all candidate variables to obtain the initial values for the nonlinear model fitting.

2) *Candidate Variables:* In the first iteration, the candidate variables consist of the source variables only. In the second iteration, candidate variables consist of attenuation and transport variables only. This is done so as to obtain an initial model with at least one source and one attenuation or transport variable. In every iteration afterwards the candidate variables can be any variable.

3) *Nonlinear Regression:* Nonlinear regression is performed by adding each candidate variable to the current model one at a time. Note that candidate variables are added according to their predetermined place in the nonlinear model (i.e. Source variables are in a linear submodel; Attenuation and transport in the exponential submodel.).

4) *Variable Selection:* The variable that results in the highest R-Squared (lowest AIC is also an option) while constrained to maintaining all variables in the model statistically significant (p-value < 0.05), is selected and added to the model. R-Squared ties beyond the thousandth decimal place are settled by the lowest p-value.

5) *Hyperparameter Optimization:* The rest of the candidate variables that differ from the selected variable by only a hyperparameter are removed from the candidate variable pool, effectively optimizing the hyperparameter value.

6) *Selection Criteria:* The new model must increase R-Squared over user-defined selection criteria such as a one percent increase. If the model passes the selection criteria, then the iterative process continues to step 2. If it does not, then the algorithm ends with the final model being the i-th minus one model since the last variable did not pass the selection criteria.

BME Estimation Framework for Space/Time Mapping Analysis

To improve estimation accuracy, we integrate the time-averaged LUR results into the Bayesian Maximum Entropy (BME) method of modern spatiotemporal geostatistics[41,42]. BME is a space/time geostatistical estimation framework grounded in epistemic principles that reduces to the space/time simple, ordinary, and universal Kriging methods as its linear limiting case when considering a limited, Gaussian, knowledge base, while also allowing the flexibility to process a wide variety of additional knowledge bases (physical laws, empirical relationships, non-Gaussian distributions, hard and soft data, etc.). We only provide the fundamental BME equations for mapping NO_3^-; the reader is referred to other works for more detailed derivations of BME equations[41,43] and LUR integration into BME[19].

Let $Z(\boldsymbol{p})$be the space/time random field (S/TRF) describing the distribution of groundwater log- NO_3^-across space and time, where $\boldsymbol{p} = (\boldsymbol{s}, t)$, $\boldsymbol{s}$ is the space coordinate and t is time. The knowledge available is organized in the general knowledge base (G-KB) about the space/time trend and variability (e.g. mean, covariance) of NO_3^-across the study domain, and the site-specific knowledge base (S-KB) corresponding to the hard and soft data $\boldsymbol{z_d}$ available at a set of specific space/time points $\boldsymbol{p}_d$.

First, we define the transformation of log- NO_3^-data $\boldsymbol{z_d}$at locations $\boldsymbol{p_d}$ as

$$\boldsymbol{x_h} = \boldsymbol{z_h} - o_Z(\boldsymbol{p_h}) \tag{1.9}$$

where $o_Z(\boldsymbol{p_h})$ may be any deterministic offset that can be mathematically calculated at any space/time coordinate $\boldsymbol{p}$. We then define $X(\boldsymbol{p})$ as a homogeneous/stationary S/TRF representing the variability and uncertainty with the transformed data $\boldsymbol{x_d}$, i.e. such that $\boldsymbol{x_d}$ is a realization of $X(\boldsymbol{p})$. Finally we let $Z(\boldsymbol{p}) = X(\boldsymbol{p}) + o_z(\boldsymbol{p})$ be the S/TRF representing groundwater log- NO_3^-. In this study, we consider two choices for $o_z(\boldsymbol{p})$**:** (1) a constant value determined by the MLE mean resulting in a purely BME model, and (2) the LUR estimate $\boldsymbol{L_z}(\boldsymbol{p_h})$ from CFN-RHO resulting in a LUR-BME model.

The G-KB for the S/TRF $X(\boldsymbol{p})$ describes its local space/time trends and dependencies. In this work, the general knowledge consists of the space/time mean trend function $m_x(\boldsymbol{p}) = E[X(\boldsymbol{p})]$, and the covariance function $C_X(\boldsymbol{p}, \boldsymbol{p}')$=$E[[X(\boldsymbol{p}) - m_x(\boldsymbol{p})][\mathrm{X}(\mathbf{p}') - m_{\boldsymbol{x}}(\boldsymbol{p}')]]$ of the S/TRF $X(\boldsymbol{p})$. The S-KB consists of hard data and soft data; with hard data, $\boldsymbol{x_h} = \boldsymbol{z_h} - \boldsymbol{L_z}(\boldsymbol{p_h})$, for data points where $\boldsymbol{z_h}$ is observed over the detection limit and soft data, $\boldsymbol{X_s}$, is at locations $\boldsymbol{p_s}$where $NO_3^-$is observed below the detection limit. Following Messier et al [19], the BME soft data for log- $NO_3^-$is modeled as a Gaussian distribution truncated above the log of the detection limit.

The overall knowledge bases considered consist of $G = \{m_x(\boldsymbol{p}), C_X(\boldsymbol{p}, \boldsymbol{p}')\}$, and $S = \{f_s(.), \boldsymbol{X_h}\}$. In this case the BME set of equations reduces to

$$f_K(x_k) = A^{-1} \int d\boldsymbol{x_s} f_G(\boldsymbol{x_h}, \boldsymbol{x_s}, x_k) f_S(\boldsymbol{x_s}) \tag{1.10}$$

where $f_K(x_k)$ is the BME posterior PDF for the offset-removed log $NO_3^-(x_k)$ at some unmonitored estimation point $\boldsymbol{p}_k$, $f_G(\boldsymbol{x_h}, \boldsymbol{x_s}, x_k)$ is the (maximum entropy) multivariate Gaussian PDF for $(\boldsymbol{x_h}, \boldsymbol{x_s}, x_k)$ with mean and variance-covariance given by G-KB, $f_S(\boldsymbol{x_s})$ is the truncated

Gaussian PDF of $\boldsymbol{X_s}$, and $A^{-1}$is a normalization constant. After the BME analysis is conducted, $o_Z(\boldsymbol{p})$ is added back to obtain log- NO_3^-concentrations.

Validation Statistics

The robustness of CFN-RHO is tested with a 10-fold cross-validation procedure. In 10-fold cross-validation data is randomly partitioned into *10* equal size subsamples. A single subsample is retained as the validation data for testing the model, and the remaining *9* subsamples are used as training data. Each of the *10* subsamples is used exactly once as the validation data**.** Similar variable selections (which may differ only by hyperparameter) for subsamples demonstrate model selection robustness.

Models are compared with a leave one-out cross-validation (LOOCV) mean squared error (MSE) and R-Squared. Spatially-smoothed/time-averaged NO_3^- and time-averaged NO_3^- models are also tested on how well they predict at the smaller observation scales. In LOOCV, each log-NO_3^-value Z_j is removed one at a time, and re-estimated using the given model based only on the remaining data. Let $Z^{*(k)}$be the re- estimate for method *k*, then $MSE^{(k)} = \frac{1}{n}\sum_{j=1}^{n}\left(Z_j^{*(k)} - Z_j\right)^2$ and the cross-validation R-Squared is $R^2(\boldsymbol{Z}, \boldsymbol{Z}^{*(k)})$**.**

Results

Nitrate Concentrations

The MLE of the statewide monitoring concentrations resulted in a geometric mean and standard deviation of the lognormal distribution of 0.62 and 14 mg/L, respectively (Figure S1.3). MLE for private wells resulted in a geometric mean and standard deviation of 0.45 and 5.1 mg/L (Figure S1.3).

Spatially-smoothed/Time-averaged Nitrate

The 25 km spatially-smoothed/time-averaged NO_3^- LUR model cross-validation results (Table 1.1) in a r^2 of 0.69 and 0.68 for monitoring and private wells, respectively, which is of similar magnitude to current literature[15]. However, as expected, the LUR model calibrated for spatially-smoothed /time-averaged NO_3^-underperforms and does progressively worse (top row, moving left to right on Table 1.1) as it predicts time-averaged $NO_3^-$and point-level NO_3^- with lower r^2 and higher RMSE. The variables selected for this model via CFN-RHO are available in the supporting information (Table S1.3).

10-fold cross-validation of spatially-smoothed/time-averaged NO_3^- LUR models was done to demonstrate the stability of CFN-RHO (Table S1.4, S1.5). All variables were selected 7 and 10 out of 10 iterations for the monitoring and private well models, respectively.

Table 1.1. Leave-one-out cross-validation statistics comparing for four estimation methods that predict spatial/temporally averaged $\boldsymbol{NO_3^-}$ concentrations, temporal averaged $\boldsymbol{NO_3^-}$ concentrations, and point-level observed $\boldsymbol{NO_3^-}$ concentrations. Note that methods were used to predict at scales more refined or equal to its calibration scale. MW = Monitoring Well model. PW= Private Well model. n = number of observations at that scale. Time averaging results in fewer observations. RMSE = Root Mean Squared Error. Units of $\boldsymbol{NO_3^-}$ concentration = mg/L.

		Predicted Value					
Method		Spatially-smoothed/Time-averaged NO_3^-		Time-averaged NO_3^-		Point-Level NO_3^-	
		MW (n=951)	PW (n=18,664)	MW (n=951)	PW (n=18,664)	MW (n=12,300)	PW (n=22,062)
Spatially-smoothed/Time-averaged LUR	r^2	0.69	0.68	0.27	0.08	0.15	0.08
	RMSE	0.895	0.293	2.23	1.19	2.40	1.27
Time-averaged LUR	r^2			0.37	0.09	0.23	0.09
	RMSE			2.08	1.19	2.28	1.27
Space/Time BME	r^2					0.70	0.25
	RMSE					1.39	1.23
Space/Time LUR-BME	r^2					**0.74**	**0.33**
	RMSE					**1.27**	**1.08**

Time-averaged Nitrate

The LUR variables selected through CFN-RHO for time-averaged NO_3^- observed at monitoring wells and private wells are shown in Table 1.2. The LUR calibrated to predict time-averaged NO_3^- obtains a r^2 of 0.37 and 0.09 for monitoring wells and private wells, respectively (Table 1.1, second row). Moreover, the LUR model predicts point-level NO_3^- with a r^2 of 0.23 and 0.09 for monitoring and private well respectively. LUR maps are available in supporting information (Figure S1.4).

Table 1.2. Nonlinear regression model variables selected via CFN-RHO and parameter estimates for time-averaged $\boldsymbol{NO_3^-}$ monitoring (left) and private well (right) models. All variables are significant with *p-value* < 0.025. Variables units: **a-** Kg- $\boldsymbol{NO_3^-}$/yr/ha, **b-** Dimensionless, **c-** 100 pigs, **d-** percent, **e-** degrees (-) Not a variable in the model.

	Monitoring Well			**Private Well**		
Variable	Variable Range	Coefficient Estimate	Standard Error	Variable Range	Coefficient Estimate	Standard Error
Constant	n/a	-3.71	0.191	n/a	-1.570	0.0382
			Source Variables			
Manure[a]	250 m	0.0759	0.0317	-	-	-
Wastewater Treatment Residuals (WTR)[b]	5 km	0.245	0.0274	-	-	-
Farm Fertilizer[a]	250 m	0.132	0.0193	250 m	0.0432	0.0025
Swine CAFO's[c]	2 km	0.117	0.0218	-	-	-
Swine Lagoons[b]	-	-	-	6 km	0.1079	0.0146
Developed Low[d]	250 m	0.112	0.0214	-	-	-
Developed	-	-	-	100 m	0.0112	7.08e-4

(All combined)[d]						
Atmospheric Deposition[a]	250 m	0.477	0.129	25 km	2.94e-11	2.53e-10
			Attenuation and Transport Variables			
Forest (All combined)[d]	2 km	-0.0064	0.00281	-	-	-
Deciduous Forest[d]	-	-	-	4 km	-0.0151	0.00127
Herbaceous Wetlands[d]	5 km	-0.531	0.079	-	-	-
Histosol[d]	25 km	-0.0427	0.0111	25 km	-0.106	0.0126
Hydrologic Soil Group D[d]	-	-	-	500 m	-0.012	0.0010
Slope[e]	25 km	-0.074	0.0261	-	-	-

10-fold cross-validation of time-averaged NO_3^- LUR models was conducted (Table S1.6, S1.7). All variables selected from the monitoring well model are selected in at least 6 iterations of the ten-fold cross-validation runs. The majority of variables in the private well model were also stable; however swine lagoons and deciduous forest were only selected 2 and 0 out of 10 times. In both models, when a variable is not selected in the 10-fold cross validation it is likely due to other variables that capture similar source, attenuation, or transport processes (i.e. Forest instead of Deciduous, Swine CAFO's instead of Swine Lagoons).

Point-Level Nitrate

We modeled the space/time covariance of the LUR offset removed log- NO_3^- S/TRF, $X(\boldsymbol{p})$, using a two-component, space/time non-separable, exponential covariance model following Messier et al[19]:

$$C_X(r,\tau) = c_1 \exp\left(-\frac{3r}{a_{r_1}}\right)\exp\left(-\frac{3\tau}{a_{\tau_1}}\right) + c_2 \exp\left(-\frac{3r}{a_{r_2}}\right)\exp\left(-\frac{3\tau}{a_{\tau_2}}\right) \quad (1.11)$$

where $c_1 = 0.67\ (log - mg/L)^2$, $a_{r_1} = 93\ m$,$a_{\tau_1} = 15\ days$, $c_2 = 3.6\ (mg/L)^2$, $a_{r_2} = 1750\ m$, $a_{\tau_2} = 15840\ days$ for monitoring wells (Figure S1.5) and a one-component, space/time exponential covariance model for private well where $c_1 = 0.76\ (log - mg/L)^2$, $a_{r_1} = 1181\ m$,$a_{\tau_1} = 8640\ days$ (Figure S1.6).

The LUR-BME model, which integrates the time-averaged LUR as the offset best predicts space/time point-level NO_3^-concentrations with a r^2 of 0.74 and 0.33 (Table 1.1) for monitoring and private wells, respectively. However, the LUR-BME predictions have a large variance at locations farther than the covariance model spatial range. Figure 1.1 maps the point-level NO_3^-concentrations estimated by LUR-BME for one day during the study period for both monitoring and private well models. These are the first results to show that there is a four-fold improvement in predicting point-level $NO_3^-$when the LUR-BME method is used in comparison to previous studies that use models for spatially-smoothed/time-averaged NO_3^-, and five percent improvement in $r^2$when integrating a LUR model into the BME framework, over purely BME. A link to a movie of LUR-BME maps is available in supporting information.

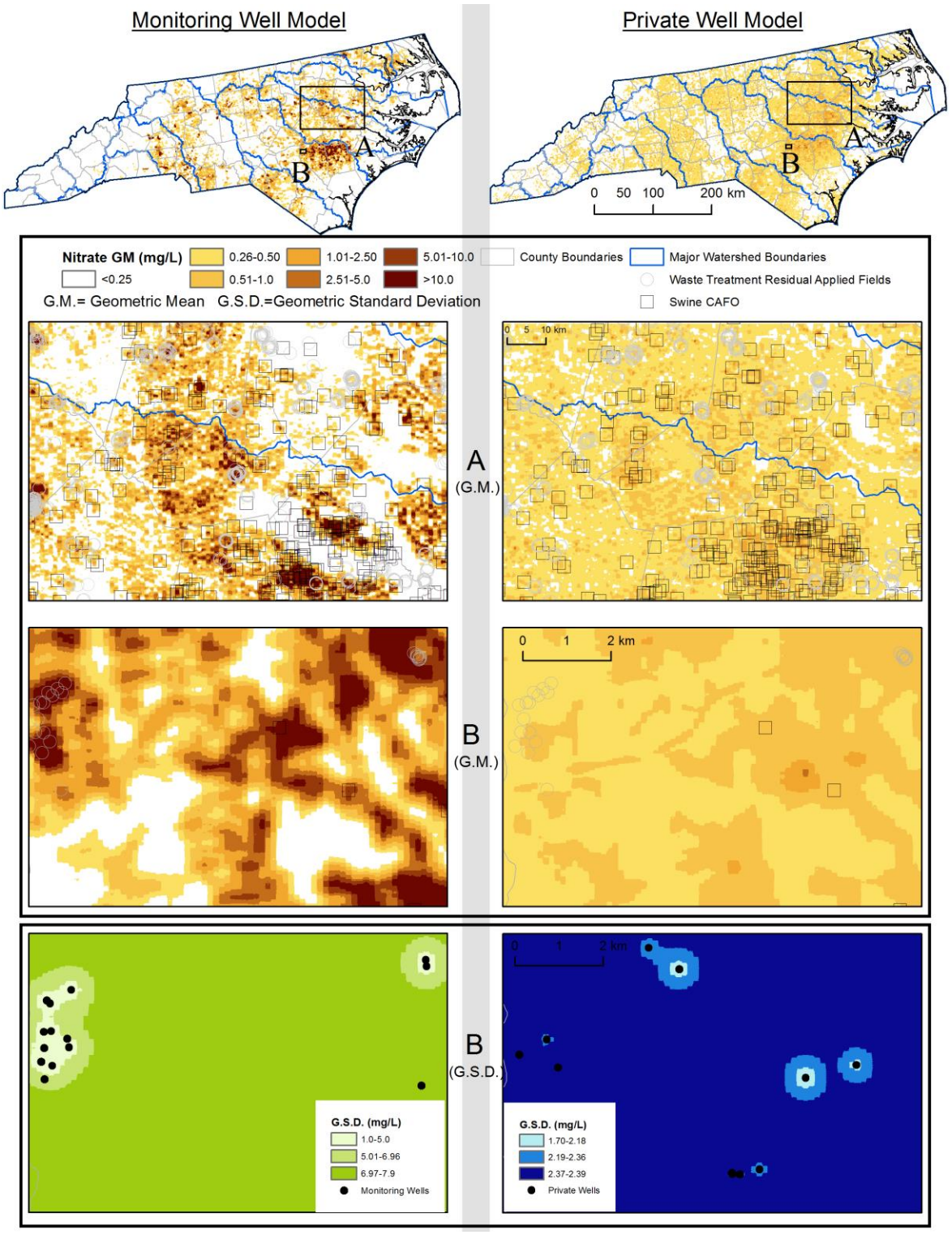


Figure 1.4. Comparison of LUR-BME results between the monitoring well (left of gray bar) model and private well (right of gray bar) model $\boldsymbol{NO_3^-}$ concentrations. The extent rectangles shows zoomed in portions of the state and are identical areas for both models. Extent (B) shows geometric mean predictions and then geometric standard deviation.

Discussion

Groundwater Nitrate Maps

This study presents a LUR model for point-level NO_3^- in North Carolina that elucidates processes affecting its local variability, and then utilizes the strengths of BME to create the first LUR-BME model of groundwater nitrate's spatial/temporal distribution including prediction uncertainty. The first major finding is the LUR-BME model for monitoring wells, assumed to represent surficial aquifers, (Figure 1.1, Movie S1) shows groundwater NO_3^- that is highly variable with many areas predicted above the current standard of 10 mg/L.

Contrarily, the private well results (Figure 1.1) depict widespread, low-level NO_3^- concentrations, which is consistent with the current physical understanding in which sources tend to pollute the surficial aquifer, but then transport over time to the deeper drinking-water supply aquifers where concentrations are lower. This finding is significant because of the studies demonstrating potential significant health effects at concentrations as low as 2.5 mg/L[4–7]. Additionally, concentrations of NO_3^- could impact ecological function since there are potential large reserves in deeper aquifers that can discharge to surface waters.[27]. The standard deviation maps (Figure 1.1) demonstrate the importance of NC-DWR and USGS monitoring wells and private well testing because areas within the spatial covariance range are well characterized, whereas those outside are less reliable.

The second major finding is the LUR-BME maps (Figure 1.1) show that groundwater NO_3^- in monitoring wells is elevated in the southeastern plains of North Carolina (Figure S1.7) due to the larger amount of NO_3^- sources and the lack of subsurface attenuation factors (Movie S2) that are present in the coastal plain region. This corroborates the findings of Nolan and Hitt[15], which also show spatially-smoothed/time-averaged NO_3^- to be the highest in the southeastern plains of North Carolina. This expands that finding with point-level results showing significant point-level variability within regional trends. Additional concerns arise since groundwater flow of the southeastern plains contributes significantly to surface water flow[27]. Our LUR-BME model can be used with surface water models to quantify the effect of groundwater NO_3^- contributing to surface water contamination.

The use of the methods in this study provide estimates at a finer resolution and down to smaller NO_3^- values than Nolan and Hitt[15], resulting in new findings. Nolan and Hitt[15] generally show greater concentrations than the LUR-BME model potentially due to their model using

significantly less training data and averaging NO_3^- over watersheds. Our LUR-BME models benefit from the large amount of monitoring (n=12,322) and private well (n=22,067) data, whereas they used 2,306 and 2,490 across the US for their shallow and drinking water models, respectively.

LUR-BME benefits from the exactitude property of BME, thus our model results are in 100% agreement at monitoring locations. Contrarily, when our observed data is compared with Nolan and Hitt[15] by grouping results according to the bins of figure 1.1, Nolan and Hitt[15] over-predicts 48% and 59% of the time for monitoring and private wells, respectively (Figure S1.8,S1.9). As a result of the finer resolution of our maps and their improved ability to predict low level NO_3^-, our results lead to a significant new finding about the extent of areas with low level contamination. Our results show private well concentrations are greater than 0.25 mg/L while monitoring well concentrations are less than 0.25 mg/L in 30.6 percent of North Carolina's area, compared to 2.6 percent for Nolan and Hitt[15] (Table S1.8,S1.9). Likewise, our results show monitoring and private wells are both above or below 0.25 mg/L at the same location in 68 percent of North Carolina, compared to 91 percent for Nolan and Hitt[15]. Hence whereas Nolan and Hitt[15] results suggest the geographical extent of the low level contamination of drinking water aquifer is limited to that of the shallow aquifer, which is consistent with downward transport of NO_3^- contamination, our LUR-BME models shows that in fact the geographical extent of the contamination of the drinking water extends over a much larger area than that of the shallow aquifer. This major new finding provides new evidence indicating that in addition to downward transport, there is also a significant outward transport of groundwater NO_3^- in the drinking water aquifer to areas outside the range of sources. This is especially significant because it indicates that the deeper aquifers are acting as a reservoir that is not only deeper, but also wider than the reservoir formed by the shallow aquifers.

LUR Variable Interpretations

Variables selected through CFN-RHO show processes influencing monitoring well and private well NO_3^- concentrations. Interpretations of regression sources parameters are based on the nonlinear model formulation: Since NO_3^- was log-transformed and the nonlinear model has multiplicative interaction, the percent increase of the geometric mean of NO_3^- is the exponential of the source coefficient multiplied by the result of the attenuation and transport terms held to their mean value. For instance, in the monitoring well model, the percent increase in the

geometric mean of NO_3^- in mg/L for every 1 kg/yr/ha of farm fertilizer is $\exp(0.132 * 0.456) = 1.06 = 5\%$ where 0.456 is the exponential of the mean attenuation and transport variables multiplied by their coefficients. For the private well model, the percent increase in the geometric mean of NO_3^- for every 1 kg/yr/ha of farm fertilizer is $\exp(0.0432 * 0.4636) = 1.02 = 2\%$. Every other source coefficient interpretation for time-averaged NO_3^- is provided in the supporting information.

Comparing variables selected between the spatially-smoothed/time-averaged NO_3^- LUR and the time-averaged NO_3^- LUR help elucidate effects the spatial scale has on groundwater NO_3^- concentrations. The variable hyperparameters selected by CFN-RHO help elucidate potential scales at which the variables affect groundwater NO_3^- concentrations. For example, the short buffer range of developed low likely captures the small size of single-family housing yards and their associated fertilizer applications. The monitoring well model WTR has an exponential decay range of 5 km. A possible explanation of this medium range is due to the volatization of NO_3^- into the air, which can then be transported over longer distances than subsurface transport mechanisms alone. Long buffer ranges for attenuation and transport variables such as percent histosol soil and mean slope represent variables with larger, regional scale effects.

The third major finding is that both wastewater treatment residuals (WTR) and swine CAFOs were selected as local sources of groundwater NO_3^- contamination, which to our knowledge have not yet been previously identified as sources in multivariable models that included regional sources. To help aide state-wide policy decisions concerning regional versus local sources, Figure 1.2 shows the elasticity of LUR predicted sources in monitoring wells, or the percent change in the geometric mean of groundwater NO_3^- within an area in response to the percent decrease in a LUR model source given all other sources remain at current levels. Farm fertilizer and atmospheric deposition result in the greatest decrease in groundwater NO_3^- state-wide (Figure 1.2A). Reducing WTR (Figure 1.2B) and swine CAFOs (Figure 1.2C) within 1 kilometer of the source leads to significant reductions in groundwater NO_3^- in the local area surrounding the sources, demonstrating the importance of sources on local area NO_3^- variability.

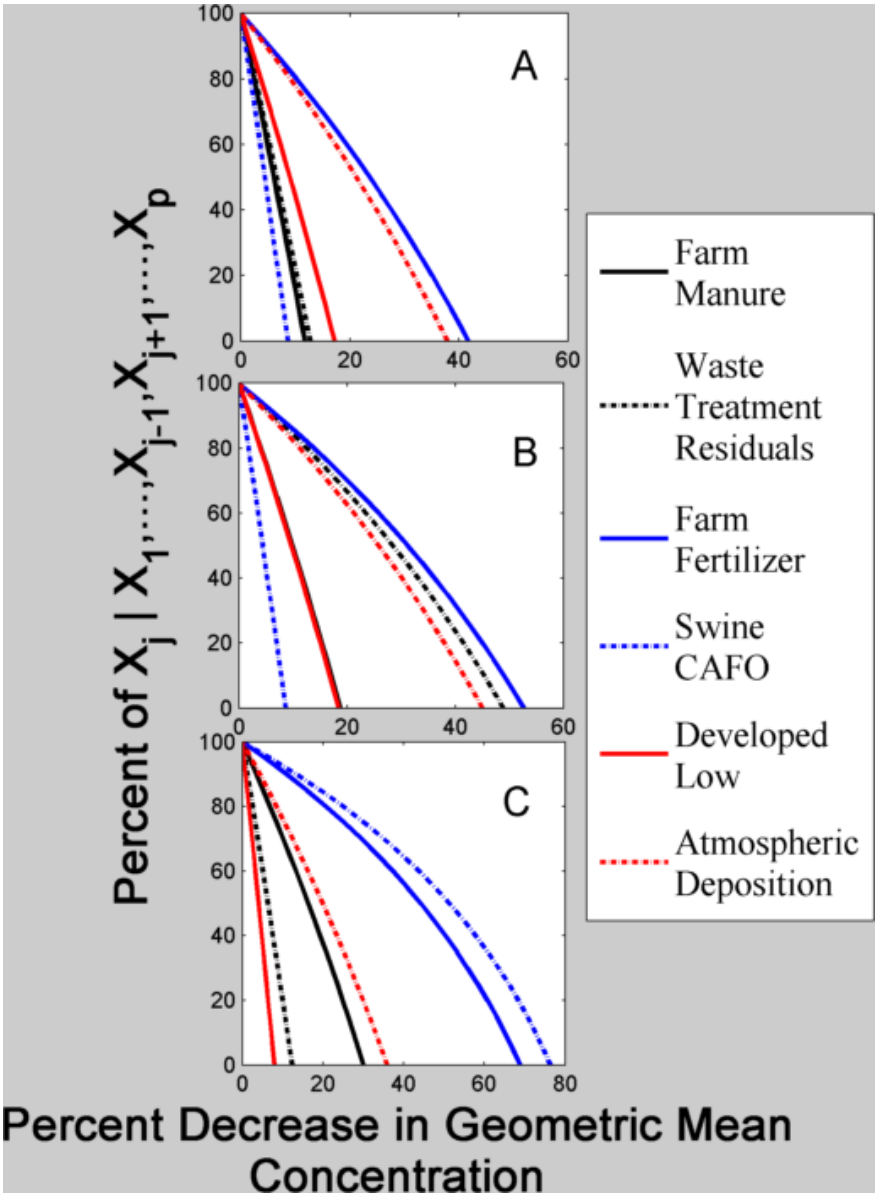


Figure 1.5. Elasticity curves for monitoring well sources. Y-axis is the percent decrease in a source and the X-axis is the percent decrease in geometric mean, for (A) State-Wide, (B) Within 1-km of Wastewater Treatment Residuals, and (C) Within 1-km of swine CAFO's.

Recommendations and Limitations

This work represents the first step in the development of modeling observed NO_3^- over large domains without averaging. In previous studies, spatial averaging is utilized because it provides results at the domain (State, Regional, or National) desired for policy making decisions and sheds light on processes influencing groundwater NO_3^-. We demonstrated that a LUR at the point-level in space is currently limited in terms of model predictive capability but when integrated into the BME framework, the improved model can estimate within the spatial

covariance range similar to LUR models for spatially-smoothed/time-averaged groundwater NO_3^- concentrations. Potential explanatory variables that can explain the remaining variability in the point-level LUR will need primary data collection. For instance, we found WTR to be a significant variable even though we just used location of fields. If records of timing and amounts of WTR applications were improved, then the temporal variability in monitoring wells near WTR application fields could be improved[44]. Similarly, a parcel-level query of farm fertilizer application practices could distinguish farms that use NO_3^- fertilizers efficiently versus farms that apply excessively or with poor timing. For private wells, the short spatial auto-correlation range may be due to differences in effectiveness of on-site wastewater treatment systems or residential fertilizer use. Additionally, we note that candidate variables not selected via CFN-RHO does not necessarily indicate they have no effect on groundwater NO_3^- concentrations in surficial or confined drinking-water aquifers of North Carolina. Many factors both statistically and physically can affect the selection such as correlation between candidate variables and local hydrogeology conditions being overwhelmed by larger scale trends. This study lacked well depth for the majority of monitoring and private wells. The monitoring and private well models clearly demonstrate a difference in concentrations based on depth, so well depth could quantify this more explicitly as opposed to categorically as done by this study. Furthermore, pumping rate information was not available for the private well data set thus the effect of local pumping could not be quantified. The USGS water use report[12] has information on domestic-use water withdrawals; however, it is at the county-scale, based on county populations, and cannot be down-scaled like the agricultural water withdrawals variable, thus it was not included as a candidate variable. Additionally, the detection limit of 1 mg/L for the private well data is high and lowering that detection limit would improve the ability of the model to delineate areas with low level contamination that may act as reservoir to surface water NO_3^- recharge. The high detection limit is also potentially responsible for the lower r^2 in the private well LUR model for time-averaged nitrate because it results in a low dependent variable variance. Predictions of the private well LUR model for time-averaged nitrate are likely biased towards the detection limit; however, the LUR-BME model for private well models likely avoids this bias due to the exactitude property along with the good spatial coverage of private well data across North Carolina. Moreover, greater uncertainty in attenuation processes in deeper aquifers is likely contributing to the lower r^2.

In conclusion, a LUR model with a novel model selection procedure can elucidate important predictors of point-level groundwater NO_3^- in North Carolina monitoring and private wells. The methods are translatable to other study areas in the United States. LUR-BME models can be used to predict spatial/temporal varying groundwater NO_3^- and provide uncertainty assessments. Further research should integrate groundwater NO_3^- results into surface water models to determine the extent of groundwater's contribution to surface water contamination. Lastly, results will be useful in identifying localities of elevated NO_3^- for increased monitoring.

Acknowledgements

This research was supported in part by funds from the NIH T32ES007018, NIOSH 2T42OH008673, and North Carolina Water Resources Research Institute (WRRI) project number 11-05-W.

Associated Content

Additional information as noted in text. This material available free of charge via the Internet at http://pubs.acs.org/.

REFERENCES

(1) US Environmental Protection Agency. Basic Information About Nitrate in Drinking Water http://water.epa.gov/drink/contaminants/basicinformation/nitrate.cfm (accessed Nov 1, 2012).

(2) Doering, O. C. I.; Galloway, J. N.; Theis, T. L.; Aneja, V.; Boyer, E.; Cassman, K. G.; Cowling, E. B.; Dickerson, R. R.; Herz, W.; Hey, D. L.; et al. *Reactive Nitrogen in the United States: An Analysis of Inputs, Flows, Consequences, and Management Options. EPA-SAB-11-013*; Washington D.C., 2011.

(3) Spalding, R. F.; Exner, M. E. Occurrence of Nitrate in Groundwater—A Review. *J. Environ. Qual.* **1993**, *22*, 392–402.

(4) Ward, M. H.; Mark, S. D.; Cantor, K. P.; Weisenburger, D. D.; Correa-Villasenor, A.; Zahm, S. H. Drinking Water Nitrate and the Risk of Non-Hodgkin ' s Lymphoma. *Epidemiology* **1996**, *7*, 465–471.

(5) Ward, M. H.; DeKok, T. M.; Levallois, P.; Brender, J.; Gulis, G.; Nolan, B. T.; VanDerslice, J. Workgroup Report: Drinking-Water Nitrate and Health—Recent Findings and Research Needs. *Environ. Health Perspect.* **2005**, *113*, 1607–1614.

(6) De Roos, A. J.; Ward, M. H.; Lynch, C. F.; Cantor, K. P. Nitrate in Public Water Supplies and the Risk of Colon and Rectum Cancers. *Epidemiology* **2003**, *14*, 640–649.

(7) Weyer, P. J.; Cerhan, J. R.; Kross, B. C.; Hallberg, G. R.; Kantamneni, J.; Breuer, G.; Jones, M. P.; Zheng, W.; Lynch, C. F. Municipal Drinking Water Nitrate Level and Cancer Risk in Older Women : The Iowa Women's Health Study. *Epidemiology* **2001**, *12*, 327–338.

(8) Paerl, H. W. Coastal eutrophication and harmful algal blooms : Importance of atmospheric deposition and groundwater as " new " nitrogen and other nutrient sources. *Limnol. Oceanogr.* **1997**, *42*, 1154–1165.

(9) Zhou, M.; Shen, Z.; Yu, R. Responses of a coastal phytoplankton community to increased nutrient input from the Changjiang (Yangtze) River. *Cont. Shelf Res.* **2008**, *28*, 1483–1489.

(10) Smith, V. H.; Tilman, G. D.; Nekola, J. C. Eutrophication: impacts of excess nutrient inputs on freshwater, marine, and terrestrial ecosystems. *Environ. Pollut.* **1999**, *100*, 179–196.

(11) USEPA. http://water.epa.gov/lawsregs/rulesregs/sdwa/index.cfm.

(12) Kenny, J. F.; Barber, N. L.; Hutson, S. S.; Linsey, K. S.; Lovelace, J. K.; Maupin, M. A. Estimated Use of Water in the United States in 2005. *USGS Circ. 1344* **2005**.

(13) Fuhrer, G. J.; Gilliom, R. J.; Hamilton, P. A.; Morace, J. L.; Nowell, L. H.; Rinella, J. F.; Stoner, J. D.; Wentz, D. A. The Quality of Our Nation's Waters: Nutrients and Pesticides. *U.S. Geol. Surv. Circ. 1225* **1999**.

(14) Daniel III, C. C.; Dahlen, P. R. Preliminary Hydrogeologic Assessment and Study Plan for a Regional Ground-Water Resource Investigation of the Blue Ridge and Piedmont Provinces of North Carolina. *U.S. Geol. Surv. Water-Resources Investig. Rep. 02-4105* **2002**.

(15) Nolan, B. T.; Hitt, K. J. Vulnerability of shallow groundwater and drinking-water wells to nitrate in the United States. *Environ. Sci. Technol.* **2006**, *40*, 7834–7840.

(16) Su, J. G.; Jerrett, M.; Beckerman, B. A distance-decay variable selection strategy for land use regression modeling of ambient air pollution exposures. *Sci. Total Environ.* **2009**, *407*, 3890–3898.

(17) Nuckols, J. R.; Beane Freeman, L. E.; Lubin, J. H.; Airola, M. S.; Baris, D.; Ayotte, J. D.; Taylor, A.; Paulu, C.; Karagas, M. R.; Colt, J.; et al. Estimating Water Supply Arsenic Levels in the New England Bladder Cancer Study. *Environ. Health Perspect.* **2011**, *1002345*.

(18) Kim, D.; Miranda, M. L.; Tootoo, J.; Bradley, P.; Gelfand, A. E. Spatial Modeling for Groundwater Arsenic Levels in North Carolina. *Environ. Sci. Technol.* **2011**, *45*, 4824–4831.

(19) Messier, K. P.; Akita, Y.; Serre, M. L. Integrating address geocoding, land use regression, and spatiotemporal geostatistical estimation for groundwater tetrachloroethylene. *Environ. Sci. Technol.* **2012**, *46*, 2772–2780.

(20) Rodríguez-Lado, L.; Sun, G.; Berg, M.; Zhang, Q.; Xue, H.; Zheng, Q.; Johnson, C. A. Groundwater arsenic contamination throughout China. *Science* **2013**, *341*, 866–868.

(21) Hoos, A. B.; Mcmahon, G. Spatial analysis of instream nitrogen loads and factors controlling nitrogen delivery to streams in the southeastern United States using spatially referenced regression on watershed attributes (SPARROW) and regional classification frameworks. *Hydrol. Process.* **2009**, *23*, 2275–2294.

(22) Howarth, R. W.; Billen, G.; Swaney, D.; Townsend, a.; Jaworski, N.; Lajtha, K.; Downing, J. a.; Elmgren, R.; Caraco, N.; Jordan, T.; et al. Regional nitrogen budgets and riverine N & P fluxes for the drainages to the North Atlantic Ocean: Natural and human influences. *Biogeochemistry* **1996**, *35*, 75–139.

(23) Smith, R. A.; Schwarz, G. E.; Alexander, R. B. Regional interpretation of water-quality monitoring data. *Water Resour. Res.* **1997**, *33*, 2781–2798.

(24) Qian, S. S. Nonlinear regression modeling of nutrient loads in streams: A Bayesian approach. *Water Resour. Res.* **2005**, *41*, 1–10.

(25) Cressie, N.; Majure, J. J. Spatio-Temporal Statistical Modeling of Livestock Waste in Streams. *J. Agric. Biol. Environ. Stat.* **1997**, *2*, 24–47.

(26) Giese, G. I.; Eimers, J. L.; Coble, R. W. Simulation of ground-water flow in the Coastal Plain system of North Carolina. *US Geol. Surv. Prof. Pap. 1404-M* **1993**, 142.

(27) Tesoriero, A. J.; Duff, J. H.; Saad, D. A.; Spahr, N. E.; Wolock, D. M. Vulnerability of Streams to Legacy Nitrate Sources. *Environ. Sci. Technol.* **2013**, *47*, 3623–3629.

(28) Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288.

(29) Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* **2005**, *67*, 301–320.

(30) Peduzzi, P. N.; Hardy, R. J.; Holford, T. R. A stepwise variable selection procedure for nonlinear regression models. *Biometrics* **1980**, *36*, 511–516.

(31) Hosmer, D. W.; Lemeshow, S. *Applied Logistic Regression*; John Wiley and Sons, 2000.

(32) Huang, C.; Townshend, J. R. G. A stepwise regression tree for nonlinear approximation: Applications to estimating subpixel land cover. *Int. J. Remote Sens.* **2003**, *24*, 75–90.

(33) Harden, S. L.; Cuffney, T. F.; Terziotti, S.; Kolb, K. R. Relation of Watershed Setting and Stream Nutrient Yields at Selected Sites in Central and Eastern North Carolina , 1997-2008. *U.S. Geol. Surv. Sci. Investig. Rep. 2013-5007* **2013**.

(34) Sanders, A. P.; Messier, K. P.; Shehee, M.; Rudo, K.; Serre, M. L.; Fry, R. C. Arsenic in North Carolina: public health implications. *Environ. Int.* **2011**, *38*, 10–16.

(35) McLay, C. D.; Dragten, R.; Sparling, G.; Selvarajah, N. Predicting groundwater nitrate concentrations in a region of mixed agricultural land use: a comparison of three approaches. *Environ. Pollut.* **2001**, *115*, 191–204.

(36) Gurdak, J. J.; Qi, S. L. Vulnerability of Recently Recharged Groundwater in Principle Aquifers of the United States To Nitrate Contamination. *Environ. Sci. Technol.* **2012**, *46*, 6004–6012.

(37) Moran, M. J.; Zogorski, J. S.; Squillace, P. J. Chlorinated solvents in groundwater of the United States. *Environ. Sci. Technol.* **2007**, *41*, 74–81.

(38) Helsel, D. R. More Than Obvious: Better Methods for Interpreting Nondetect Data. *Environ. Sci. Technol.* **2005**, *39*, 419A–423A.

(39) Pradhan, S. S.; Hoover, M. T.; Austin, R. E.; Devine, H. A. *Potential Nitrogen Contributions from On-site Wastewater Treatment Systems to North Carolina ' s River Basins and Sub-basins*; Raleigh, North Carolina, 2007.

(40) Beven, K. J.; Kirkby, M. J. A physically based, variable contributing area model of basin hydrology. *Hydrol. Sci.* **1979**, *24*, 43–69.

(41) Christakos, G. A Bayesian/maximum-entropy view to the spatial estimation problem. *Math. Geol.* **1990**, *22*, 763–777.

(42) Serre, M. L.; Christakos, G. Modern geostatistics: computational BME analysis in the light of uncertain physical knowledge - the Equus Beds study. *Stoch. Environ. Res. Risk Assess.* **1999**, *13*, 1–26.

(43) Christakos, G.; Bogaert, P.; Serre, M. L. *Temporal GIS: Advanced Function for Field-Based Applications*; Springer: New York, NY, 2002.

(44) Keil, A.; Wing, S.; Lowman, A. Suitability of Public Records for Evaludating Health Effects of Treated Sewage Sludge in North Carolina. *NC Med J.* **2011**, *72*, 98–104.

Supporting Information for Chapter 1

Nitrate Variability in Groundwater of North Carolina using Monitoring and Private Well Data Models

Kyle P. Messier[†]*, Evan Kane*[‡]*, Rick Bolich*[‡]*, Marc L. Serre*[†]***

Authors' Affiliation:

† Department of Environmental Science and Engineering, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC

‡ North Carolina Department of Environment and Natural Resources, Division of Water Resources

***Corresponding Author:**

Marc L. Serre

Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, University of North Carolina, 1303 Michael Hooker Research Center, Chapel Hill, NC 27599

Phone: (919) 966-7014 Fax: (919) 966-7911

The Supporting Information includes 27 pages, 9 tables, 9 figures, and 2 movie links.

Spatial Explanatory Variables

1) Nitrate Mass in Fertilizer, Manure, and Atmospheric Deposition. Estimates of nitrate were based on USGS estimates of nitrate mass in farm fertilizer, non-farm fertilizer, manure, and atmospheric deposition. The estimates are based on county-level estimates compiled from fertilizer sales, census of agriculture, and population estimates following the methods outlined in Ruddy et al.[1], and employed by Hoos and McMahon[2] for the analysis of nitrogen loads in streams using spatially referenced regression on watershed attributes (SPARROW).

Nitrate mass estimates in kilograms per year per county was obtained from Ruddy et al[1] and averaged over all of the available years to obtain an average mass per year per county estimate. Similar to Hoos and McMahon[2], in order to more accurately represent the spatial distribution of the county-level data, nitrate farm fertilizer and manure estimates were distributed to only agricultural land according to the 2006 National Land Cover Database [3]. The non-farm fertilizer was distributed to the developed, forest, shrub, and grassland land cover classes. The atmospheric deposition was distributed evenly across each county. The total amount of nitrate mass per area for each county was divided by the number of 30-meter cells within each county that was portioned mass estimates resulting in variables that represent the average amount of nitrate mass input (from the respective source) per year per square-meter, which is then multiplied by 900 square-meters to obtain nitrate mass per year. Following the creation of nitrate mass variables, we calculate the mean nitrate mass per year per hectare from each source (*l=Farm Mass, Non-Farm Mass, Manure, or Atmospheric deposition)* as:

$$NM_i^{(l)}(\lambda_l) = \frac{1}{\pi\lambda^2}\sum_{j=1}^{n_i(\lambda_l)} M_j^{(l)} \qquad \text{(S1.12)}$$

where $NM_i^{(l)}(\lambda_l)$ is the mean nitrate mass per year per hectare of type (*l*) within a radius λ_l of nitrate point *i*, $M_j^{(l)}$ is the estimated nitrate mass (kg/year) of type *l* for the j^{th} pixel described above surrounding nitrate point *i* , $\pi\lambda^2$ is the area of the circular buffer, and $n_i(\lambda_l)$ is the number of pixels within the circular buffer of radius λ_l around nitrate point *i*. Area units are converted from square meters to hectares, which is more common in the agricultural field.

2) Point Source Variables. Following Messier et al.[4], we calculate the sum of exponentially decaying contribution from various potential nitrate point sources including wastewater

treatment residuals (WTR) application fields[5], swine farms, swine waste lagoons, cattle farms, chicken farms, and wastewater treatment plants (WWTP). Equation 2 shows the general form of the point source variables,

$$PS_i^{(l)}(\lambda_l) = \sum_{j=1}^{n_l} C_{0j}^{(l)} \exp\left(-3 * \frac{D_{ij}}{\lambda_l}\right) \quad \text{(S1.13)}$$

where $PS_i^{(l)}(\lambda_l)$ is the sum of exponentially decaying contribution from point sources type (l) at nitrate point i, n_l is the total number of point sources of type (l), D_{ij} is the distance between the j-th point source of type (l) and the nitrate point i, C_{0j} is a proxy for the initial nitrate concentration at the point source if available, or equal to 1 otherwise, and λ_l is the exponential decay range corresponding to the distance it takes for nitrate released by source of type (l) to be reduced by 95%. WWTP initial values are based on the design capacity of the plant; cattle, chicken, and swine farms are weighted based on the number of animals; and the other point source variables do not have information available to provide reasonable estimates of the initial concentration.

3) On-Site Wastewater Treatment. On-site wastewater treatment, or septic tanks, variables are created following the methods of Pradhan et al[6] with adjustments for our variables' circular buffers as opposed to watershed polygons. The 1990 US census was the last census to collect information on the method of wastewater treatment used in residential homes, which was obtained at the census block group level as the number of septic or other on-site wastewater treatment systems (i.e. latrine, straight pipe) per census block group. We calculated the estimated septic system density as follows:

$$SD_i(\lambda) = \frac{\sum_{j=1}^{n_i(\lambda)} \xi_j^{(\lambda)}}{\pi\lambda^2} \quad \text{(S1.14)}$$

where $SD_i(\lambda)$ is the septic system density (#/mi^2) around nitrate point i within circular buffer λ, $n_i(\lambda)$ is the total number of census block groups within circular buffer λ, $\xi_j^{(\lambda)}$ is the number of septic systems in the overlapping area between census block j and the circle created by radius λ

assuming a constant density of septic tanks in each census block, and $\pi\lambda^2$ equals the area of the circular buffer created with radius λ.

The average nitrate loading from septic system is

$$SN_i(\lambda) = \sum_{j=1}^{n_i(\lambda)} PD_j * a_{j\lambda} * p_j * 10 \quad \text{(S1.15)}$$

where $SN_i(\lambda)$ is the septic nitrate (lb/yr) around nitrate point *i* circular buffer λ, $n_i(\lambda)$ is the total number of census block groups within circular buffer λ, PD_j is the population density (people/mi^2) in census block group *j*, $a_{j\lambda}$ is the area of overlap between census block group *j* and λ, p_j is the proportion of people (dimensionless) in census block *j* that are on septic systems, and the result is multiplied by 10 lb/person-year based on the worst case-scenario that the amount of nitrate septic influent is estimated at 10 pounds per person per year [6].

4) Population density. Population density represents a surrogate variable associated with non-farm nitrate inputs and is calculated for each circular buffer using the 2000 census population data at the block level and assumes population is evenly distributed over each block.

5) National Land Cover Database. We construct explanatory variables based on the National Land Cover Database (NLCD) satellite imagery file that characterizes land cover types at 30 meter resolution. We create variables for every NLCD land cover type and aggregated land cover type that represent attenuation variables including deciduous forest, evergreen forest, mixed forest, herbaceous wetlands, and woody wetlands . For a NLCD variable (*l*) of interest we calculate

$$LC_i^{(l)}(\lambda_l) = \frac{1}{n_i(\lambda_l)} \sum_{j=1}^{n_i(\lambda_l)} I_j^{(l)} \quad \text{(S1.16)}$$

where $LC_i^{(l)}(\lambda_l)$ is the percent of land cover of type (*l*) within a radius λ_l of nitrate point *i*, $I_j^{(l)}$ is an indicator variable equal to 1 if the j^{th} pixel surrounding nitrate point *i* is of type *l*, and zero otherwise, and $n_i(\lambda_l)$ is the number of pixels within the circular buffer of radius λ_l around nitrate point *i*.

6) *Slope and Topographic Wetness Index.* Slope and Topographic Wetness Index (TWI) [7] are variables that represent possible attenuation and transport variables and are calculated from a digital elevation raster. Slope is calculated as the average gradient between adjacent cells within a circular buffer centered on each well. TWI expresses the potential wetness in soils due to topography and is commonly used in watershed scale hydrological models [7,8] and as a predictor variable for groundwater contaminants [9]. The mean TWI within a circular buffer is calculated as

$$TWI_i(\lambda) = \frac{1}{n_i(\lambda)} \sum_{j=1}^{n_i(\lambda)} \ln\left(\frac{F_{Aj}}{\tan(\beta_j)}\right) \quad \text{(S1.17)}$$

where F_{Aj} is the j-th flow accumulation calculated from a D8 flow algorithm, and β_j is the j-th pixel slope, and $n_i\ (\lambda)$is the number of pixels that are within radius λ around nitrate point *i*.

7) *Soil variables.* Soil based variables are calculated as the average of the given soil characteristic within a circular buffer. We use the multilayer soil characteristics dataset for the conterminous United States (CONUS-SOIL), which contains soil estimates of pH, permeability, hydrologic soil groups, available water capacity, and depth to bedrock [10]. Data on histosol soil type, a soil group that contains large amounts of organic matter in the upper profile, was obtained directly from the supporting information of Nolan and Hitt[11].

8) *USGS withdrawals.* Similar to Nolan and Hitt[11], we calculate the average water withdrawals from groundwater, surface water, and the sum of groundwater and surface water. Water withdrawal rates per county [12] are distributed evenly over each county, which is then used to calculate the average water withdrawal within a circular buffer.

Model Coefficient Interpretations

Interpretations of regression sources parameters are based on the nonlinear model formulation: Since nitrate was log-transformed and the nonlinear model has multiplicative interaction, the percent increase of the geometric mean of nitrate is the exponential of the source coefficient multiplied by the result of the attenuation and transport terms held to their mean value. Below is the derivation of this interpretation:

In matrix format, let us write an equation for the log of the nitrate with the equation form in this paper, with the attenuation and transport term simplified into one exponential term.

$$Ln\ (N) = X\beta \exp(Z\gamma)$$

For simplicity, let's reduce it to one source and one attenuation/transport variable.

$$Ln\ (N) = \beta_1 X_1 \exp(\gamma_1 Z_1)$$

Let us write another equation that represents a one unit increase in source X_1.

$$Ln\ (N_2) = \beta_1 (X_1 + 1)\exp(\gamma_1 Z_1)$$

For clarity, rename $N = N_1$ and evaluate the attenuation/transport term at the mean values, leading to a constant value. We have two equations:

$$\begin{cases} Ln\ (N_1) = \beta_1 X_1 K \\ Ln\ (N_2) = \beta_1 (X_1 + 1) K \end{cases}$$

Subtract the equations and simplify

$$Ln(N_1) - Ln(N_2) = -\beta_1 K$$

$$-B_1 K = Ln\left(\frac{N_1}{N_2}\right)$$

$$\beta_1 K = Ln\left(\frac{N_2}{N_1}\right)$$

$$\exp(\beta_1 K) = N_2/N_1$$

Using the derived formula the model source interpretations for the monitoring well model are as follows:

1) The percent increase in the geometric mean of nitrate in mg/L for every 1 kg/yr/ha of farm manure while other sources and attenuation/transport is constant is $\exp(0.0759 * 0.456) = 1.04 = 4\%$.

2) The percent increase in the geometric mean of nitrate in mg/L for every 1 unit of wastewater treatment residuals while other sources and attenuation/transport is constant is $\exp(0.245 * 0.456) = 1.12 = 12\%$.

3) The percent increase in the geometric mean of nitrate in mg/L for every 1 kg/yr/ha of farm fertilizer while other sources and attenuation/transport is constant is $\exp(0.132 * 0.456) = 1.06 = 6\%$.

4) The percent increase in the geometric mean of nitrate in mg/L for every 100 pigs in swine CAFO's while other sources and attenuation/transport is constant is $\exp(0.117 * 0.456) = 1.06 = 6\%$.

5) The percent increase in the geometric mean of nitrate in mg/L for every 1 percent increase in developed low land while other sources and attenuation/transport is constant is $\exp(0.112 * 0.456) = 1.05 = 5\%$.

6) The percent increase in the geometric mean of nitrate in mg/L for every 1 kg/yr/ha of nitrate in atmospheric deposition while other sources and attenuation/transport is constant is $\exp(0.447 * 0.456) = 1.23 = 23\%$.

For private wells:

1) The percent increase in the geometric mean of nitrate in mg/L for every 1 kg/yr/ha of farm fertilizer is while other sources and attenuation/transport is constant $\exp(0.0432 * 0.4636) = 1.02 = 2\%$.

2) The percent increase in the geometric mean of nitrate in mg/L for every 10 percent increase in developed land while other sources and attenuation/transport is constant is $\exp(0.0112 * 0.4636 * 10) = 1.05 = 5\%$.

3) The percent increase in the geometric mean of nitrate in mg/L for every 1 unit of swine lagoons while other sources and attenuation/transport is constant is $\exp(0.1079 * 0.4636) = 1.05 = 5\%$.

4) The percent increase in the geometric mean of nitrate in mg/L for every 100 kg/yr/ha of nitrate in atmospheric deposition while other sources and attenuation/transport is constant is $\exp(2.9e-11 * 0.4636 * 100) = 1.02 = 0.0000000014\%$. This seemingly negligible increase is due to the fact that the hyperparameter is 25km, thus the increase in atmospheric deposition in widely distributed.

Tables

Table S1.3. Groundwater Nitrate Data Source Basic Information.

Data Source	Median (mg/L)	Mean (mg/L)	Unique Wells	Space/Time Samples	Year Range	Percent Detected
NC-DWR	1.30	4.61	366	11,004	1980-2011	79.7
USGS	0.10	6.14	585	1,318	1990-2012	61.4
Private Well	0.62	1.66	18,664	22,067	1990-2011	30.6

Table S1.4. Spatial explanatory variable model category. The candidate variables are listed according to their category in the groundwater $\boldsymbol{NO_3^-}$ model. Details on how each variable calculated is presented in the previous section of the supporting information.

	Sources	Attenuation	Transport
Variable Names	Farm Fertilizer; Non-Farm Fertilizer; Manure; Nitrate Atmospheric Deposition; Points Source: WWTP, Cattle Farms, Poultry Farms, Swine Farms, Swine Lagoons, Waste Treatment Residuals (WTR); On-Site Wastewater Treatment input; On-Site Wastewater treatment density;	National Landcover Database: Deciduous, Evergreen, Mixed Forest, Forest All, Grassland, Woody Wetlands, Herbaceous Wetlands, Wetlands All; Histosol Soils	Soil Permeability; Depth to Bedrock; pH; Hydrologic Soil Groups: A,B,C,D; Available Water Capacity; Water Withdrawals: Groundwater, Surface Water, Total; Topographic Wetness Index; Mean Slope

	National Landcover Database: Developed Open, Developed Low, Developed Medium, Developed High, Developed All, Pasture/Hay, Crops, Agriculture combined		

Table S1.5. Nonlinear regression model variables selected via CFN-RHO and parameter estimates for spatially-smoothed/time-averaged NO_3^- monitoring (left) and private well (right) models. All variables are significant with *p-value* < 0.025. Variables units: **a-** Kg- NO_3^-/yr/ha, **b-** Dimensionless, **c-** 100 pigs, **d-** percent, **e-**cubic meters per second. (-) Not a variable in the model.

25 KM Spatially Smoothed/Temporally Averaged Nitrate

	Monitoring Well			Private Well		
Variable	Variable Range	Coefficient Estimate	Standard Error	Variable Range	Coefficient Estimate	Standard Error
Constant	**n/a**	-3.71	0.191	n/a	-1.570	0.0382
			Source Variables			
Wastewater Treatement Residuals (WTR)[b]	40 km	0.0235	0.0056	-	-	-
Farm Fertilizer[a]	25 km	4.67e-9	8.0e-10	25 km	7.2e-10	3.5e-11
Swine Lagoons[b]	-	-	-	35 km	0.0385	0.0016
Atmospheric	25 km	3.07e-8	4.8e-9	25 km	8.49e-9	1.4e-10

Deposition[a]						
Wastewater Treatment Plant	25 km	0.0132	0.0003	-	-	-
			Attenuation and Transport Variables			
Deciduous Forest[d]	25 km	-0.0416	0.0026	25 km	-0.0312	5.5e-4
Mixed Forest	-	-	-	25 km	-0.0395	0.0021
Herbaceous Wetlands[d]	25 km	-0.7042	0.0649	25 km	-0.1757	0.0112
Histosol[d]	25 km	-0.0482	0.0076	25 km	-0.0924	0.0037
Hydrologic Soil Group D[d]	25 km	-0.013	0.0019	25 km	-0.0271	5.7e-4
Hydrologic Soil Group C[d]	25 km	-0.0123	0.0027	-	-	-
GWW[e]	-	-	-	25 km	-1.8014	0.0448

Table S1.6. The number of times each variable in the full spatially-smoothed/time-averaged LUR model for monitoring wells was selected in the ten-fold cross-validation runs.

Variable	Number out of 10 the variable was picked in 10 fold cross-validation
Farm Mass	10
NADP	7
WWTP	9

WTR	**10**
Deciduous	**10**
Herbaceous Wetlands	**10**
HSG-C	**7**
HSG-D	**8**
Histosols	**10**

Table S1.7. The number of times each variable in the full spatially-smoothed/time-averaged LUR model for private wells was selected in the ten-fold cross-validation runs.

Variable	**Number out of 10 the variable was picked in 10 fold cross-validation**
Farm Mass	10
Atmospheric Deposition	10
Swine Lagoons	10
HSG D	10
Deciduous	10
Herbaceous Wetlands	10
GWW	10
Histosol	10

Table S1.8. The number of times each variable in the full time-averaged LUR model for monitoring wells was selected in the ten-fold cross-validation runs.

Variable	Number out of 10 the variable was picked in 10 fold cross-validation
Manure	6
WTR	10
Farm Fertilizer	10
Swine CAFO's	10
Developed Low	7
Atmospheric Deposition	7
Forest	7
Herbaceous Wetlands	10
Histosol	8
Slope	7

Table S1.9. The number of times each variable in the full time-averaged LUR model for private wells was selected in the ten-fold cross-validation runs.

Variable	Number out of 10 the variable was picked in 10 fold cross-validation
Farm Fertilizer	10

Developed	**10**
Swine Lagoons	**2**
Atmospheric Deposition	**7**
Histosol	**7**
HSG D	**10**
Deciduous	**0**

Table S1.10. 2 x 2 table showing the percent of area in North Carolina as predicted by this study's LUR-BME model to be (I) below 0.25 mg/L for both monitoring and private wells, (II) above 0.25 mg/L for monitoring wells and below 0.25 for private wells, (III) below 0.25 mg/L for monitoring wells and above 0.25 mg/L for private wells, and (IV) above 0.25 mg/L for both monitoring and private wells.

		Monitoring Well	
		<0.25 mg/L	**>=0.25 mg/L**
Private Well	**<0.25mg/L**	***I*** **43.2**	***II*** **1.4**
	>=0.25mg/L	***III*** **30.6**	***IV*** **24.8**

Table S1.11. 2 x 2 table showing the percent of area in North Carolina as predicted by thi GWAVA models (Nolan and Hitt, 2006) to be (I) below 0.25 mg/L for both monitoring and private wells, (II) above 0.25 mg/L for monitoring wells and below 0.25 for private wells, (III) below 0.25 mg/L for monitoring wells and above 0.25 mg/L for private wells, and (IV) above 0.25 mg/L for both monitoring and private wells.

		Shallow Groundwater	
		<0.25 mg/L	**>=0.25 mg/L**
Drinking Water	**<0.25mg/L**	***I*** **25.4**	***II*** **6.0**
	>=0.25mg/L	***III*** **2.6**	***IV*** **66.0**

Figures

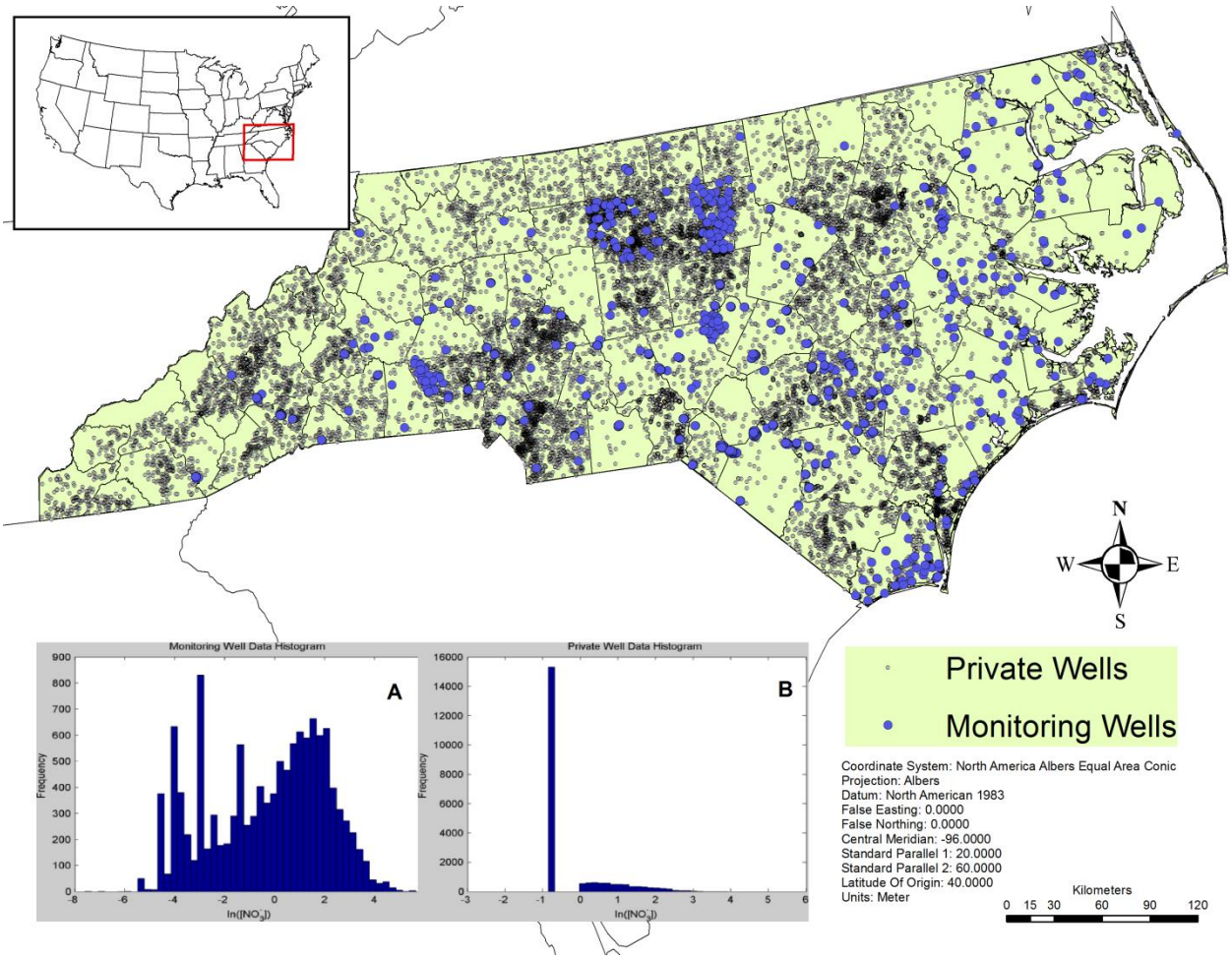


Figure S1.6. North Carolina study area with private well and monitoring well nitrate databases. The convex hull of monitoring and private wells covers 88 and 99.5 percent of North Carolina, respectively. A) Frequency histogram of the log-nitrate concentration for monitoring well data. B) Frequency histogram of the log-nitrate concentration for private well data.

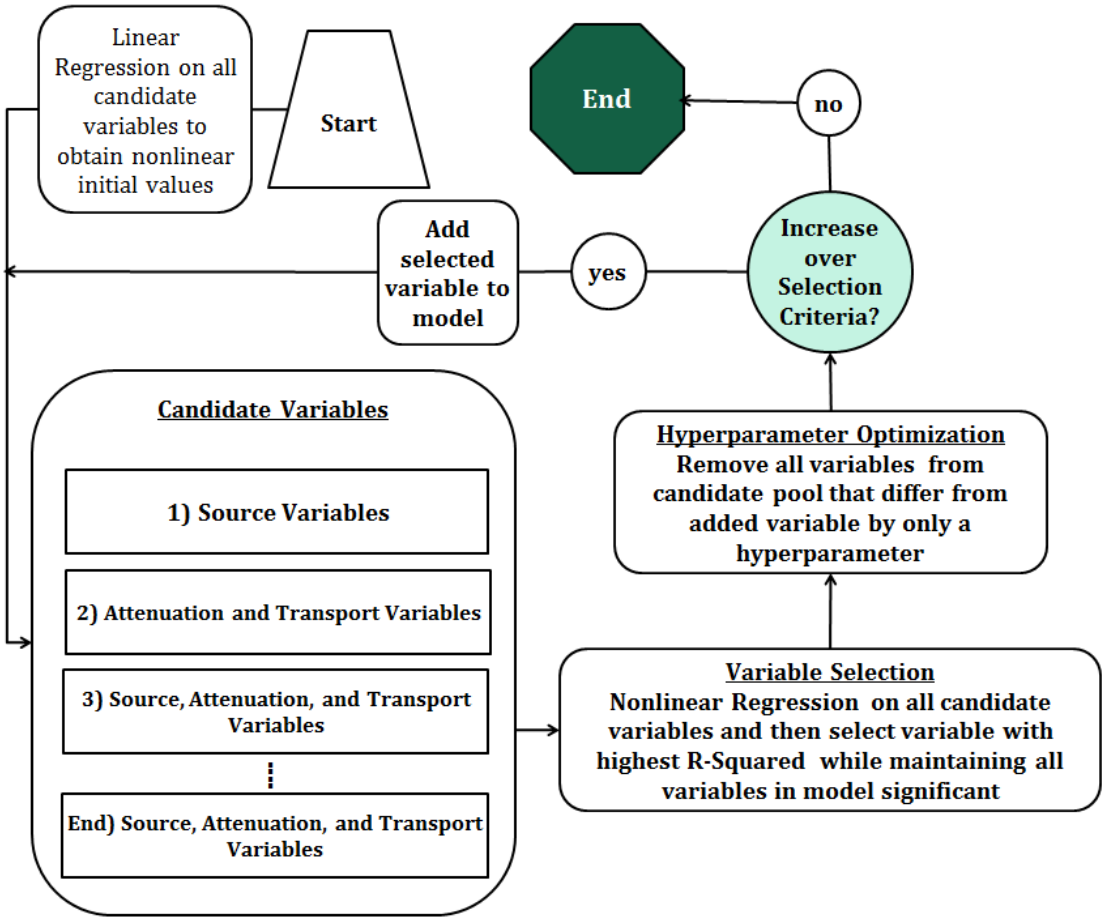


Figure S1.7. Flow diagram of the constrained forward nonlinear and hyperparameter optimization model selection procedure.

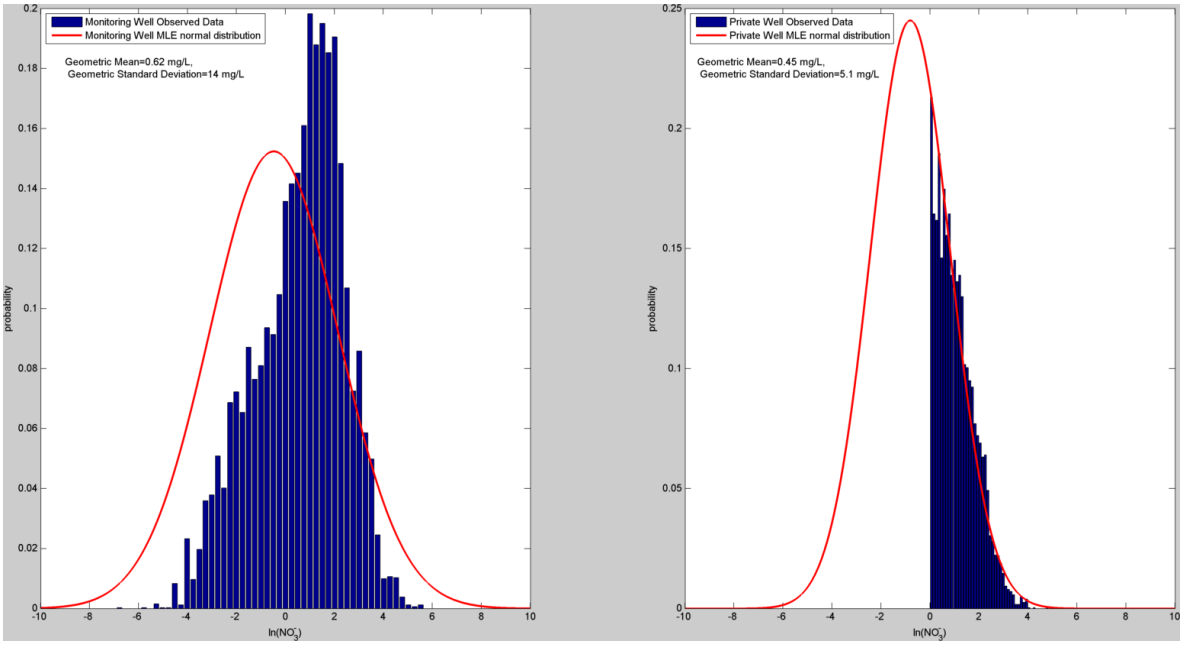


Figure S1.8. Left) Histogram (blue) of monitoring well data only observed above the detection limit, log-transformed. The fitted normal distribution (red) based on the maximum likelihood estimation method accounting for nondetects and their detection limits. Right) Histogram (blue) of private well data only observed above the detection limit, log-transformed. The fitted normal distribution accounting for nondetects (red).

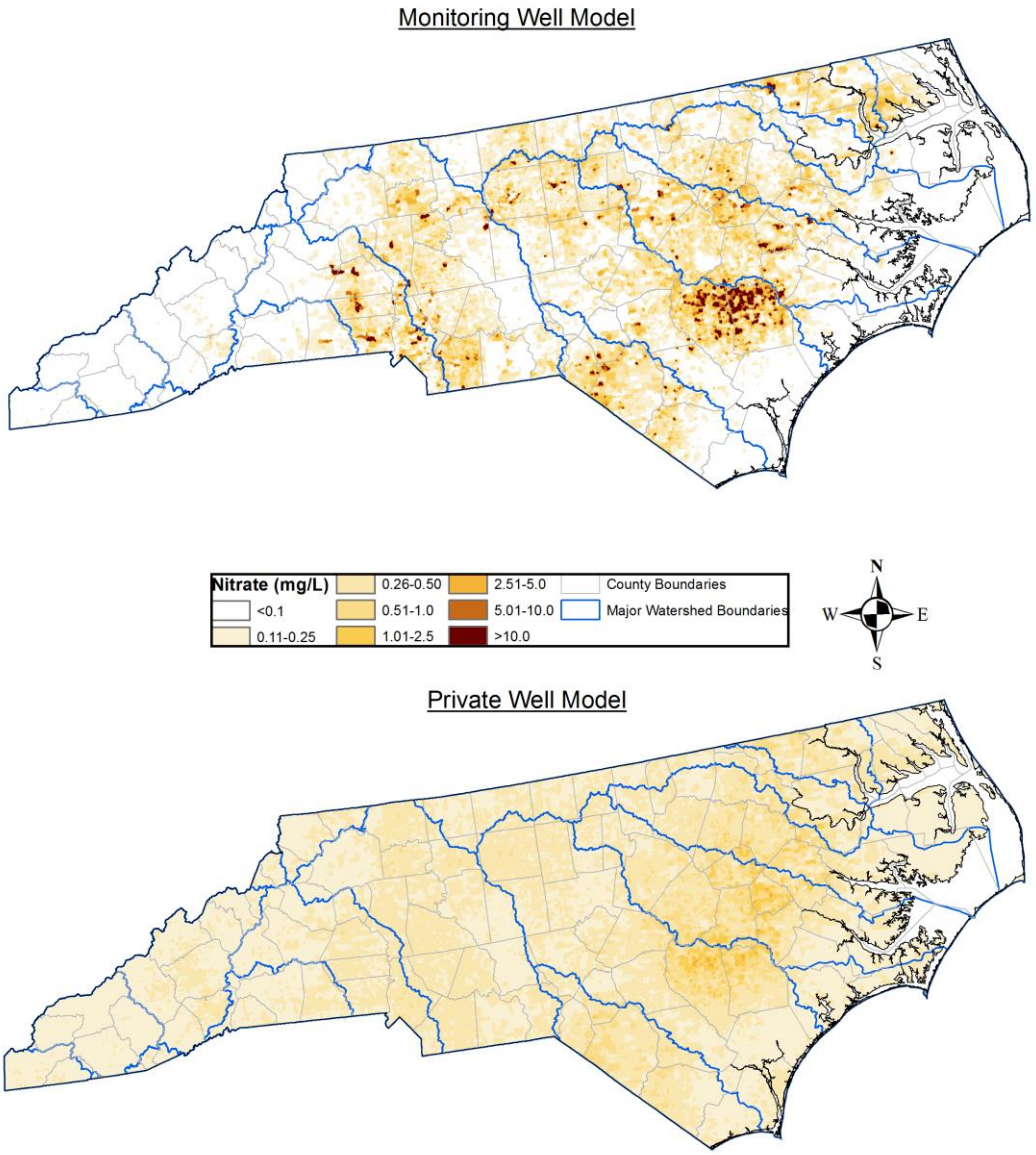


Figure S1.9. Land Use Regression results from the Constrained Forward Nonlinear Regression and Hyperparameter Optimization procedure for the monitoring and private well models. There are significant areas of predicted nitrate above 10 mg/L in the southeastern plains region for the monitoring wells. This area also has relatively widespread contamination above 1 mg/L in the private wells. Prediction variance should be used in conjunction with results at unmonitored locations.

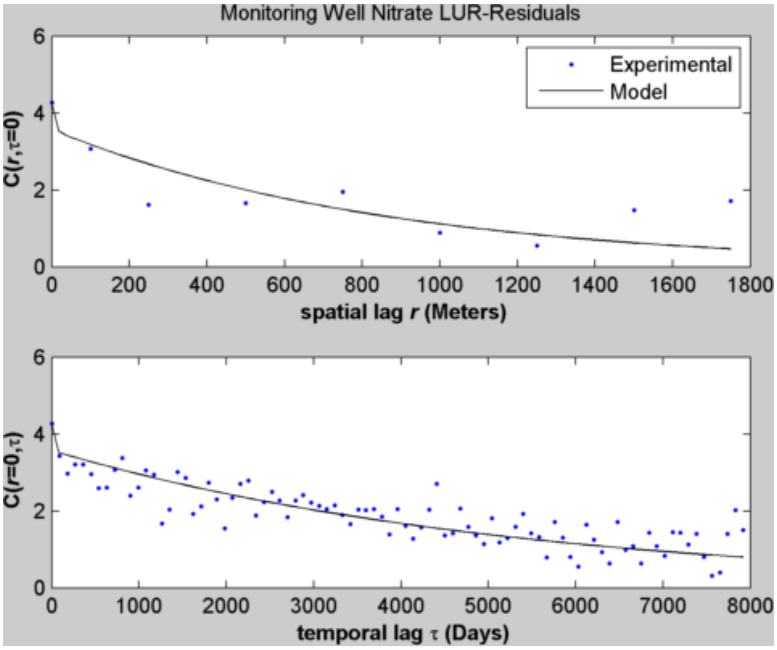


Figure S1.10. Monitoring well nitrate LUR residual experimental and modeled spatial (top) and temporal (bottom) covariance. The model is fit based on a least-squared fit with weights equal to the experimental covariance at the lag times the square root of the number of pairs used to calculate the covariance.

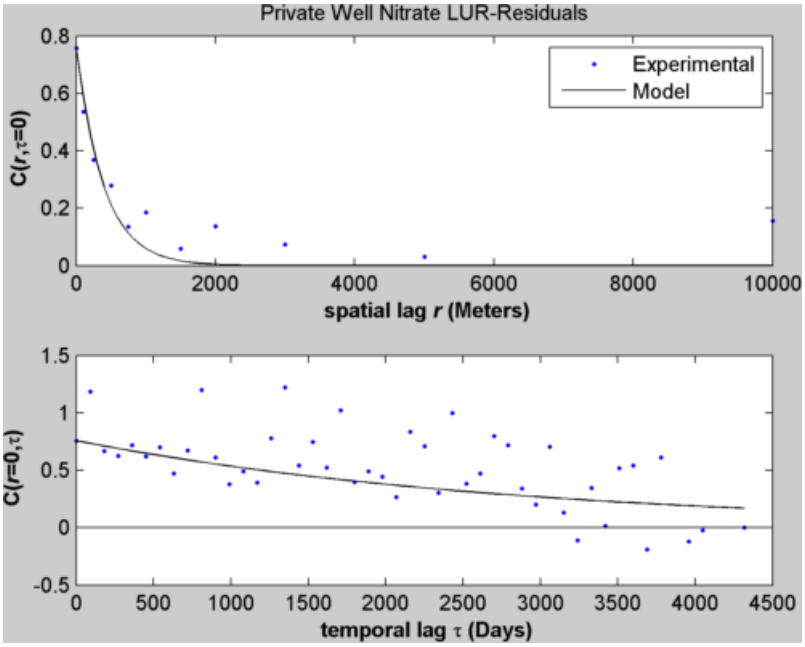


Figure S1.11. Private well nitrate LUR residual experimental and modeled covariance.

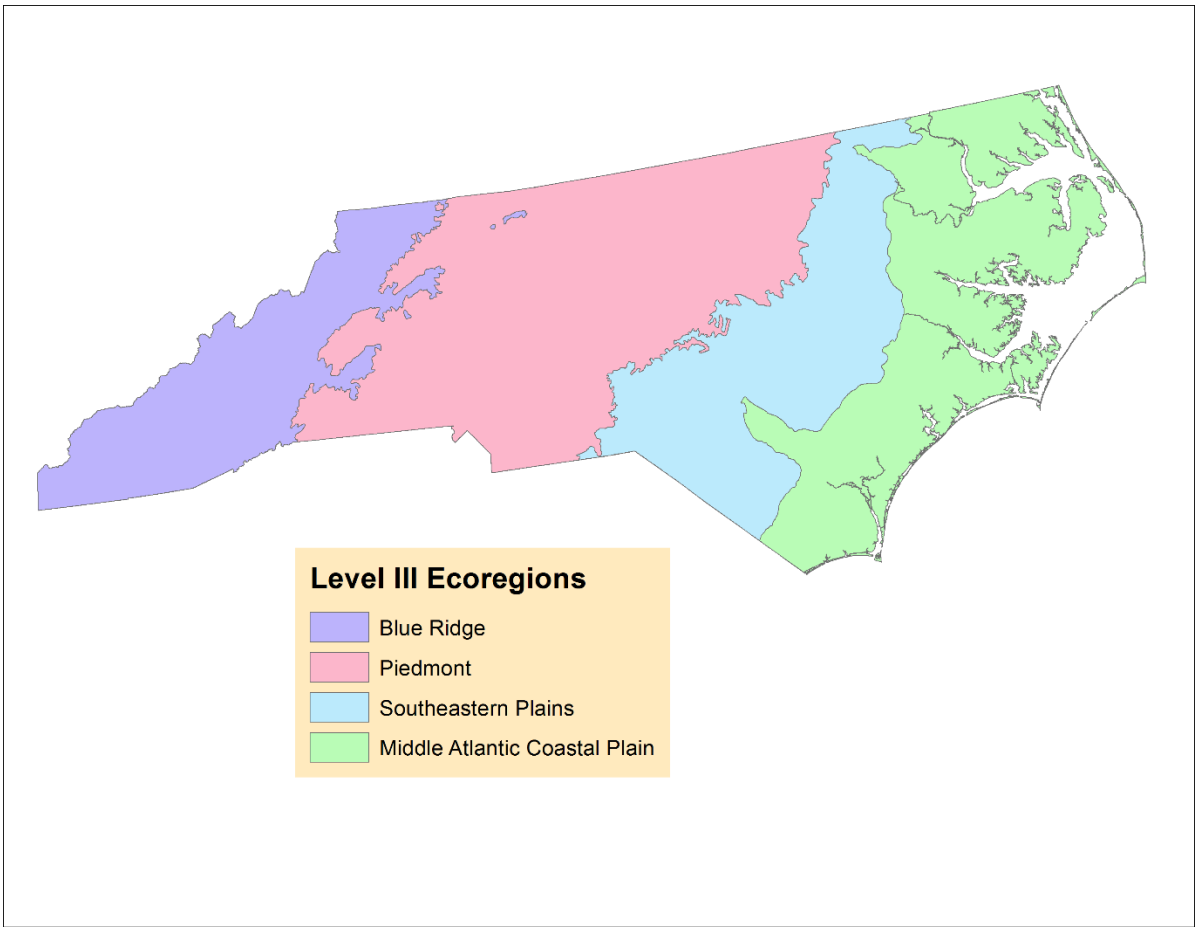
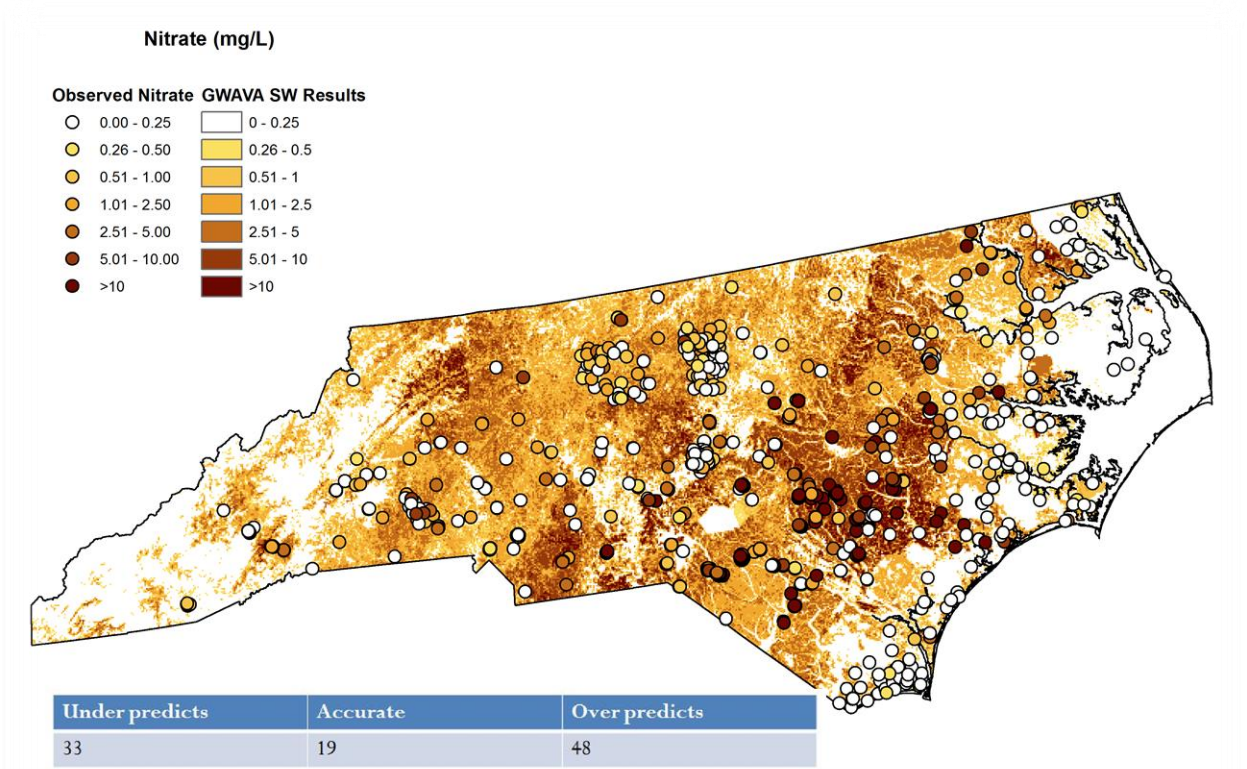


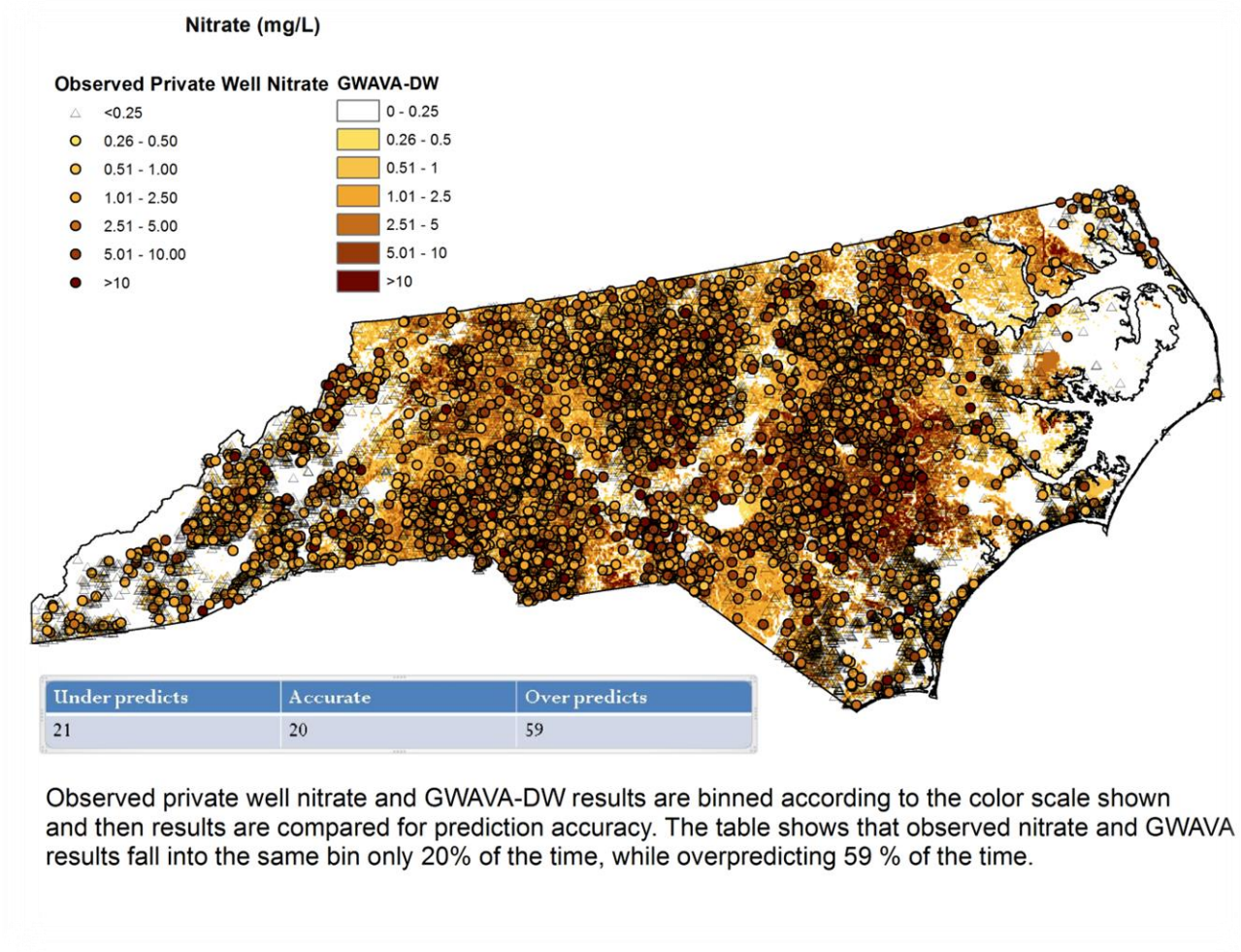
Figure S1.12. Level III Ecoregions in North Carolina defined by the US Environmental Protection Agency.



Under predicts	Accurate	Over predicts
33	19	48

Observed nitrate and GWAVA-SW results are binned according to the color scale shown and then results are compared for prediction accuracy. The table shows that observed nitrate and GWAVA results fall into the same bin only 19% of the time, while overpredicting almost half the time.

Figure S1.13. Observed monitoring well nitrate from this study overlaid with the GWAVA-SW model results.



Under predicts	Accurate	Over predicts
21	20	59

Observed private well nitrate and GWAVA-DW results are binned according to the color scale shown and then results are compared for prediction accuracy. The table shows that observed nitrate and GWAVA results fall into the same bin only 20% of the time, while overpredicting 59 % of the time.

Figure S1.14. Observed private well nitrate from this study overlaid with the GWAVA-DW model results.

Movies

Movie S1: A movie showing the LUR-BME estimates for multiple days across the study time period is available for viewing and download at http://www.unc.edu/depts/case/BMElab/studies/KM_NO3_NC/

Movie S2: A movie showing the explanatory variables for the monitoring well LUR model is available for viewing and download at http://www.unc.edu/depts/case/BMElab/studies/KM_NO3_NC/

REFERENCES

(1) Ruddy, B. C.; Lorenz, D. L.; Mueller, D. K. County-Level Estimates of Nutrient Inputs to the Land Surface of the Conterminous United States , 1982 – 2001 Scientific Investigations Report 2006 – 5012 County-Level Estimates of Nutrient Inputs to the Land Surface of the Conterminous United States , 19. **2006**, 1982–2001.

(2) Hoos, A. B.; Mcmahon, G. Spatial analysis of instream nitrogen loads and factors controlling nitrogen delivery to streams in the southeastern United States using spatially referenced regression on watershed attributes (SPARROW) and regional classification frameworks. *Hydrol. Process.* **2009**, *23*, 2275–2294.

(3) Fry, J.; Xian, G.; Jin, S.; Dewitz, J.; Homer, C.; Yang, L.; Barnes, C.; Herold, N.; Wickham, J. Completion of the 2006 National Land Cover Database for the Conterminous United States. *Photogramm. Eng. Remote Sensing* **2011**, *77*, 858–864.

(4) Messier, K. P.; Akita, Y.; Serre, M. L. Integrating address geocoding, land use regression, and spatiotemporal geostatistical estimation for groundwater tetrachloroethylene. *Environ. Sci. Technol.* **2012**, *46*, 2772–2780.

(5) Keil, A.; Wing, S.; Lowman, A. Suitability of Public Records for Evaludating Health Effects of Treated Sewage Sludge in North Carolina. *NC Med J.* **2011**, *72*, 98–104.

(6) Pradhan, S. S.; Hoover, M. T.; Austin, R. E.; Devine, H. A. *Potential Nitrogen Contributions from On-site Wastewater Treatment Systems to North Carolina ' s River Basins and Sub-basins*; Raleigh, North Carolina, 2007.

(7) Beven, K. J.; Kirkby, M. J. A physically based, variable contributing area model of basin hydrology. *Hydrol. Sci.* **1979**, *24*, 43–69.

(8) Tague, C. L.; Band, L. E. RHESSys: Regional Hydro-Ecologic Simulation System—An Object-Oriented Approach to Spatially Distributed Modeling of Carbon, Water, and Nutrient Cycling. *Earth Interact.* **2004**, *8*, 1–42.

(9) Rodríguez-Lado, L.; Sun, G.; Berg, M.; Zhang, Q.; Xue, H.; Zheng, Q.; Johnson, C. A. Groundwater arsenic contamination throughout China. *Science* **2013**, *341*, 866–868.

(10) Miller, D.; White, R. A Conterminous United States Multi-Layer Soil Characteristics Data Set for Regional Climate and Hydrology Modeling. *Earth Interact.* **1998**, *2*.

(11) Nolan, B. T.; Hitt, K. J. Vulnerability of shallow groundwater and drinking-water wells to nitrate in the United States. *Environ. Sci. Technol.* **2006**, *40*, 7834–7840.

(12) Kenny, J. F.; Barber, N. L.; Hutson, S. S.; Linsey, K. S.; Lovelace, J. K.; Maupin, M. A. Estimated Use of Water in the United States in 2005. *USGS Circ. 1344* **2005**.

CHAPTER 2

Estimation of Groundwater Radon in North Carolina using Land Use Regression and Bayesian Maximum Entropy

Kyle P. Messier[†], *Ted Campbell*[‡], *Phil Bradley*[§], *Marc L. Serre*[†]*

Authors' Affiliation:

†Department of Environmental Science and Engineering, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC 27599, United States

‡ North Carolina Department of Environment and Natural Resources, Division of Water Resources, 2090 U.S. 70 Highway, Swannanoa, NC 28778, United States

§ North Carolina Geological Survey, 1620 Mail Service Center, Raleigh, North Carolina 27699, United States

***Corresponding Author:**

Marc L. Serre

Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, University of North Carolina, 1303 Michael Hooker Research Center, Chapel Hill, NC 27599

Phone: (919) 966-7014 Fax: (919) 966-7911

Acknowledgements:

This research was supported in part by funds from the NIH T32ES007018, NIOSH 2T42OH008673, and North Carolina Water Resources Research Institute (WRRI) project number 11-05-W.

Abstract

Radon (^{222}Rn) is a naturally occurring chemically inert, colorless, and odorless radioactive gas produced from the decay of uranium (^{238}U), which is found in rocks and soils worldwide. Exposure to ^{222}Rn is likely the second leading cause of lung cancer after cigarette smoking via inhalation; however, exposure through untreated groundwater is also a contributing factor to both inhalation and ingestion routes. A land use regression (LUR) model for groundwater ^{222}Rn with anisotropic geological and ^{238}U based explanatory variables is developed, which helps elucidate the factors contributing to elevated ^{222}Rn across North Carolina. The LUR is also integrated into the Bayesian Maximum Entropy (BME) geostatistical framework to produce a point-level LUR-BME model of groundwater ^{222}Rn across North Carolina including prediction uncertainty. The LUR-BME model of groundwater ^{222}Rn results in a leave-one out cross-validation r^2 of 0.46 (Pearson correlation coefficient= 0.68), effectively predicting within the spatial covariance range. Results show ^{222}Rn concentration differences between Intrusive Felsic geological formations is likely due to sediment ^{238}U concentrations.

Introduction

Radon (^{222}Rn) is a naturally occurring chemically inert, colorless, and odorless radioactive gas [1] produced from the decay of uranium (^{238}U), which is found in rocks and soils worldwide. Outdoor air ^{222}Rn levels are generally very low; however, when ^{222}Rn enters a residential home, its concentration can increase to levels that may lead to adverse health effects [1]. There is vast literature supporting the conclusion that exposures via inhalation of indoor air contaminated with radon lead to a significant increased risk of lung cancer morbidity in both never-smokers and smokers [2–7] . Exposure to ^{222}Rn is likely the second leading cause of lung cancer after smoking in the US[4,8,9]. Important routes of inhalation exposure result from ^{222}Rn gas directly escaping from soil and rock and accumulating in the indoor environment; however, ^{222}Rn can also degas from untreated groundwater used for showering, dishwashing, and clothes washing resulting in exposures in direct vicinity to the breathing zone [10,11].

^{222}Rn in groundwater is not only a concern because of its contribution to indoor air ^{222}Rn, but also due to the direct ingestion of drinking water with elevated ^{222}Rn. There is evidence that exposure to ^{222}Rn through drinking water and indoor air can lead to stomach cancer [8,12]; however, this human health endpoint is understudied compared to lung cancer and there is not a consensus among the literature [4].

The association between groundwater ^{222}Rn and underlying geological formations has been shown in many previous studies. Brutsaert et al. (1981) found positive associations between ^{222}Rn and granites, metamorphic rocks, and other chemical parameters in Maine, USA through graphical and tabular comparison of measured values. Further solidifying this relationship, Yang et al. [14] showed increased risk for elevated ^{222}Rn within a 5 km distance to granitic intrusions in Maine, USA using the non-parametric Kruskal-Wallis one-way analysis of variance (ANOVA). Likewise, associations between elevated ^{222}Rn and granites and granitic gneisses have been shown in North Carolina[8,15]. Prediction of groundwater ^{222}Rn on medium to large area scales (>10^0 km) has been reasonably successful with Kriging models [16] and multivariate statistics [17]; however they do not account for physical processes affecting its distribution such as geochemistry and geology interaction.

Previous studies have also attempted to find associations and make predictions of groundwater ^{222}Rn based on ^{238}U and other hydrogeochemical parameters such as alkalinity and conductivity. Yang et al. [14] observed weak, but positive correlations at intermediate scales (10^0-

10^1 km) between ^{238}U and ^{222}Rn in granitic bedrock aquifers of Maine, USA. Salih et al. [18] used Co-Kriging with ^{238}U as the secondary variable to map groundwater ^{222}Rn in southeast Sweden, which produced good predictions at unmonitored locations, but had weak correlation with ^{238}U (R^2 <0.1).

About 25 percent of the Piedmont and mountains physiographic provinces of North Carolina are underlain with rocks commonly associated with elevated ^{222}Rn in water, namely felsic intrusive rocks such as granites and granitic gneisses. Through water sampling Campbell et al.(2011) have found 19 counties in North Carolina that are particularly susceptible to elevated radon in water. In this study, we use the samples from Campbell et al. (2011) plus geocoded samples from private well sources and USGS to model the groundwater ^{222}Rn concentrations across North Carolina..

Several counties in western North Carolina are classified as EPA Zone 1 counties, with predicted indoor air ^{222}Rn concentrations above the action level of 4 picocuries per liter (pCi/L). Over 90 percent of wells sampled in that region exceed the EPA's proposed Maximum Contaminant Level of 300 pCi/L and a large number exceeded the alternate MCL of 4000 pCi/L [8]. Since monitoring ^{222}Rn concentration is not mandatory for private well owners [19], elucidating the spatial distribution of radon across the state is indispensable to inform the public about exposure to waterborne ^{222}Rn. Furthermore, since North Carolina has on average 1/3 of each county population relying on untreated groundwater as drinking water [20], quantifying potential exposures is important since the population potentially exposed in North Carolina significantly higher than the United States average.

Land use regression (LUR)[21–26] modeling is a proven method that complements monitoring programs and provides effective means for water quality exposure assessments. The Bayesian Maximum Entropy (BME) method of modern spatiotemporal geostatistics has also been shown to successfully estimate groundwater quality contaminants [25,27,28]. An advantage of BME over purely spatial linear geostatistical approaches is its ability to quantify spatial and temporal variability which is then used in the estimation process at unmonitored locations. BME, like all geostatistical methods, is data driven and can only provide reliable estimates within the vicinity of measured values. However, BME utilizes Bayesian epistemic knowledge blending to combine multiple sources of data, which has been successfully demonstrated with incorporation

of deterministic mean trend functions, such as a LUR model, into BME for groundwater contaminants (Messier et al., 2014, 2012).

The objectives of this study are to: 1) Develop a linear anisotropic LUR model for point-level groundwater ^{222}Rn in North Carolina, 2) Integrate the LUR model into BME to produce the first model for point-level groundwater ^{222}Rn that fully quantifies its distribution with a mean or median and error variance, 3) Elucidate and develop hypotheses about geological and hydrogeochemical factors controlling its distribution. To these ends, we create groundwater ^{222}Rn explanatory variables based on the recent published geological and accompanying GIS information [29] and ^{238}U data [30]. Results are of interest to many parties including: 1) Agencies that regulate drinking-water sources or that monitor health outcomes from ingestion of drinking-water, 2) Agencies that monitor ^{222}Rn and provide remediation options to homeowners with increased risk of elevated radon, and 3) Geologists and hydrogeologists interested in environmental and human health applications of geological surveys.

Methods

Radon Data Sources

Groundwater ^{222}Rn data (Figure 2.1) are obtained from three data sources, which are detailed as follows:

The North Carolina Department of Environment and Natural Resources (NC-DENR) Division of Water Resources (NC-DWR) has sampled and analyzed groundwater for ^{222}Rn where levels are suspected to be elevated. This resulted in 655 samples of groundwater ^{222}Rn and their known spatial location. Samples were collected by NC DENR personnel from a plumbing fixture as close to the wellhead as possible, usually from the wellhead itself. The sample was collected after the pump had been operating for at least 20 minutes to ensure the water was not from a stagnant water column. Samples were collected using a special procedure to prevent aeration of the ^{222}Rn. Specifically, 60 milliliter glass vials were carefully submerged, filled, and sealed inside a 2 liter plastic beaker that had been filled with well water under laminar flow. The samples were then put on ice to maintain a cool temperature and shipped to a certified laboratory overnight. Most ^{222}Rnsamples were analyzed using the analytical E-Perm ion electret de-emanation procedure [31]; a smaller number were analyzed using Standard Method 7500-Rn procedure [32].

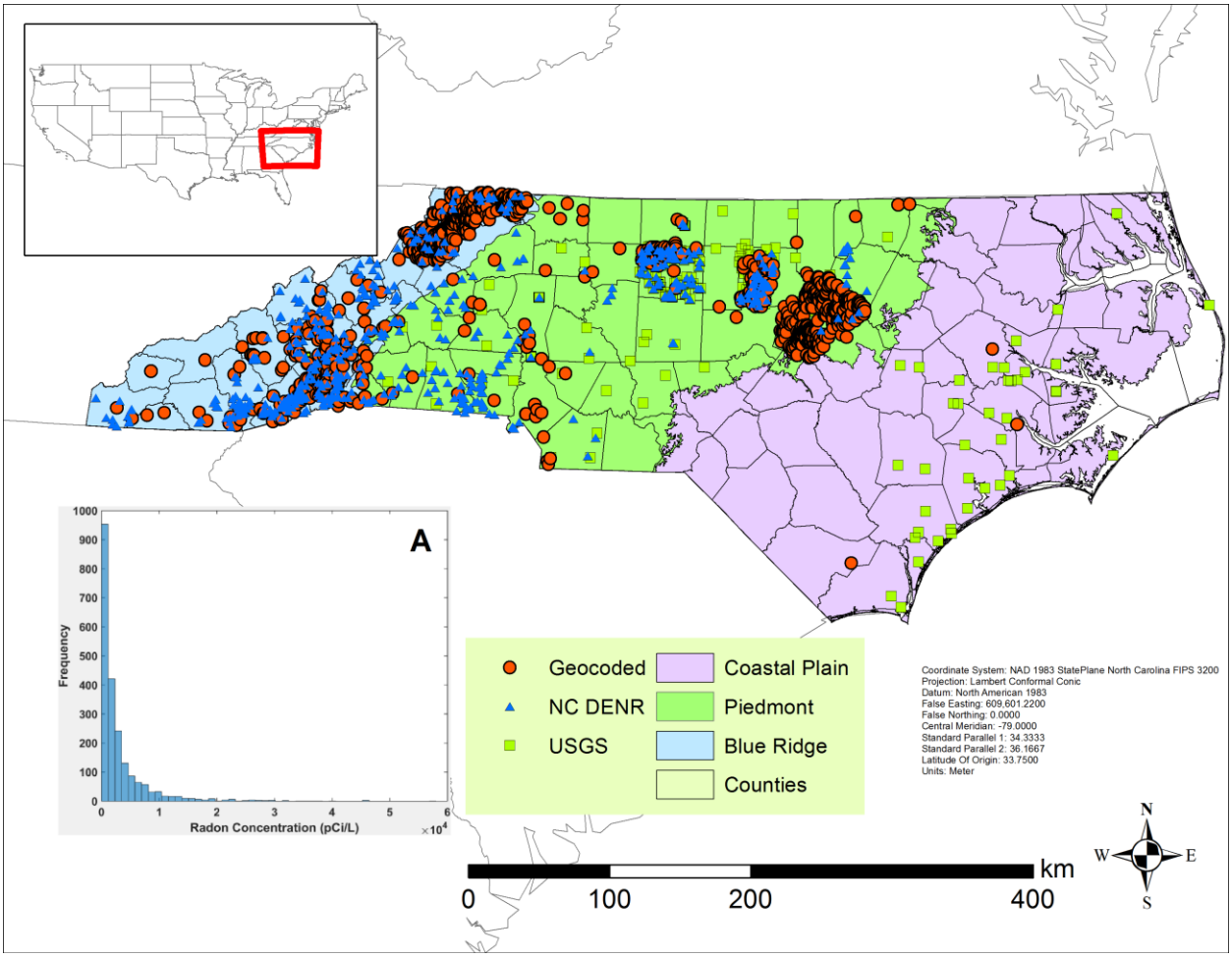


Figure 2.15. Radon data source spatial distribution detailed by its source. The 3 physiographic provinces of North Carolina are detailed by color: Coastal Plain is Light Pink, Piedmont is green, and Blue Ridge is light blue. A) Frequency histogram of the radon data. Note its lognormal distribution.

The second source is USGS data obtained through the National Water Information System (USGS), which yielded 297 groundwater unfiltered ^{222}Rn measurements (USGS parameter code 82303) . Details of the USGS sampling procedure can be obtained through USGS directly.

The last dataset of groundwater ^{222}Rncomes from private well data collected by private companies and used with permission. These data were address geocoded using the same process as outlined in Messier et al. [25]. The private company samples were analyzed using the Standard Method 7500-Rn procedure. Private home owners receive kits provided by the companies

contracted to analyze the dissolved ^{222}Rn concentrations. The kits contain detailed instructions on how to sample, store, and ship according to EPA approved methods.

Spatial Explanatory Variables

Spatial explanatory variables representing the underlying geology are calculated prior to model development. For a given geology feature, the corresponding geological variable is calculated as the percentage of that geological feature within an elliptical buffer centered on each radon measurement. Each geological variable is characterized by a set of *ellipse hyperparameters* and its *geological classification scale*, as follow:

i) Ellipse hyperparameters. The ellipse buffer used to calculate the percent of a given geological feature captures the anisotropy and spatial range of the corresponding geological formation of interest. A given ellipse is defined using a set $\mathbf{\Lambda} = (\lambda_1, \lambda_2, \phi)$ of three ellipse hyperpameters which are the major and minor ellipse buffer radii λ_1 and λ_2, respectively, and the angle ϕ of ellipse rotation with respect to the horizontal axis. Each variable is calculated with multiple hyperparameter values since these are unknown a priori. In the final model selection process a maximum of one ellipse hyperparameter set $\mathbf{\Lambda}$ is allowed to be selected for each geological variable to avoid multicollinearity and effectively optimize the hyperparameters. The ellipse axis lengths included in this study are 1000, 2500, 5000, 75000, 10000meters. The ellipse angles of rotation included are 0, 45, 90, and 135 degrees.

ii) Geological Classification Scale. Geological features are defined at 3 different geological spatial scales, which are natural to lithological descriptions of geology and allow the model to distinguish between large area and small area effects of geology on groundwater ^{222}Rn. The most general and largest area scale is referred to as *General Geological Descriptions* (subsequently referred to as *General*) and this includes descriptions such as intrusive felsic, intrusive mafic, and orthogneiss. These large area scale features are subdivided into intermediate scale geologic descriptions referred to as *Lithotectonic Element* (subsequently referred to as *Element*), which are themselves subdivided into the most detailed geologic descriptions referred to as *Units.* Maps of the geological classification at each scale are available in the supplemental data. These geological classifications are based on the underlying geology provided by Hibbard et al [29]. The provided GIS attributes by Hibbard et al (2006) were enhanced with North Carolina-centric names based on North Carolina Geological Survey information for interpretability; however, the actual extent of each geological feature remains unchanged.

For a given ellipsoid and geological feature, we define the *Geology Percent variable*, as well as several corresponding *Geology and Uranium variables*, as follow:

1) Geology Percent variable. The percent $G^{(l)}(\boldsymbol{s};\boldsymbol{\Lambda})$ of geological feature (*l*) within an ellipse ($\boldsymbol{s}$,$\boldsymbol{\Lambda}$) centered at spatial location $\boldsymbol{s}$ and with hyperparameter set $\boldsymbol{\Lambda}$ is calculated as:

$$G^{(l)}(\boldsymbol{s};\boldsymbol{\Lambda}) = \frac{1}{n_i(\boldsymbol{\Lambda})} \sum_{j=1}^{n_i(\boldsymbol{\Lambda})} I_j^{(l)}(\boldsymbol{s};\boldsymbol{\Lambda}) \tag{2.18}$$

where $I_j^{(l)}$ is an indicator representing the presence/absence of geological feature (*l*) at the *j*-th pixel in the ellipse, and n_i is the total number of pixels within the ellipse.

2) *Geology and Uranium variables.* For each stand-alone geological percent variable we define several corresponding geology and uranium variables that combine geological information with uranium information obtained from the National Uranium Resource Evaluation (NURE) Hydrogeochemical and Stream Sediment Reconnaissance [30] data. The geology and uranium variable $H_i^{(l)}(\boldsymbol{s};\boldsymbol{\Lambda})$ is calculated for geological feature (*l*) within the ellipse ($\boldsymbol{s}$,$\boldsymbol{\Lambda}$) as the product of the geological percent variable $G_i^{(l)}(\boldsymbol{s};\boldsymbol{\Lambda})$ times the average uranium, or normalized uranium concentration, in that geological feature within the ellipse, i.e.

$$H^{(l)}(\boldsymbol{s};\boldsymbol{\Lambda}) = G^{(l)}(\boldsymbol{s};\boldsymbol{\Lambda}) \left(\frac{1}{m_i(\boldsymbol{s};\boldsymbol{\Lambda})} \sum_{j=1}^{m_i(\boldsymbol{s};\boldsymbol{\Lambda})} U_j^{(l)}(\boldsymbol{s};\boldsymbol{\Lambda}) \right) \tag{2.19}$$

where $U_j^{(l)}(\boldsymbol{s};\boldsymbol{\Lambda})$ is the concentration of ^{238}U in the groundwater or stream sediment, or the ^{238}U concentration normalized by alkalinity, or ^{238}U normalized by conductivity, at the *j*-th grid cell in the ellipse $(\boldsymbol{s};\boldsymbol{\Lambda})$ that contains geology (*l*), and $m_i(\boldsymbol{s};\boldsymbol{\Lambda})$ is the number of raster data grid cells for geology (*l*) in the ellipse $(\boldsymbol{s};\boldsymbol{\Lambda})$. ^{238}U normalized variables are included as potential variables because they help remove ^{238}U anomalies [34,35] and stream sediment variables are included because ^{238}U solubility and groundwater flow makes it tend to accumulate near streams. The details of equations 2.1 and 2.2 are aided by figure 2.2, which shows an example of the denominator for an equation 1, $n(\boldsymbol{\Lambda})$, and a geological formation of interest, $m(\boldsymbol{s};\boldsymbol{\Lambda})$.

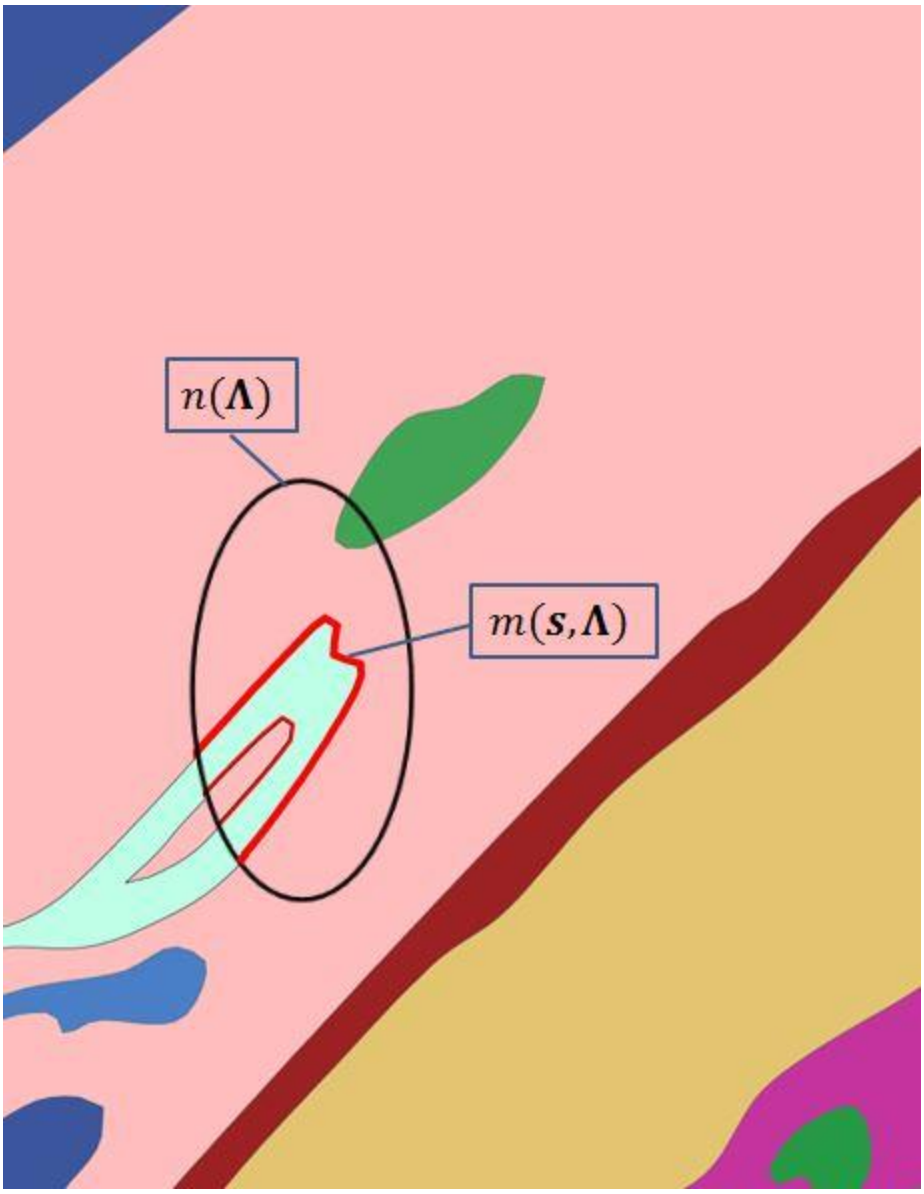


Figure 2.16. $\boldsymbol{n}(\boldsymbol{\Lambda})$ is the number of cells in ellipse $(\boldsymbol{s};\boldsymbol{\Lambda})$ (Black line) located at $\boldsymbol{s}$ and with hyperparameters $\boldsymbol{\Lambda}$, and $\boldsymbol{m}(\boldsymbol{s},\boldsymbol{\Lambda})$ is the number of cells in the geology of interest (Red line) in the ellipse. In this example, the light blue represents the geological formation of interest, and $\boldsymbol{m}(\boldsymbol{s},\boldsymbol{\Lambda})$ corresponds to the area outlined in red.

Land Use Regression and Model Selection

We implement a linear land use regression (LUR) model for ^{222}Rn concentration as follows:

$$Y_i = \beta_0 + \sum_{l=1}^{L} \beta_l X_i^{(l)}(\boldsymbol{\Lambda}_l) + \varepsilon_i \tag{2.20}$$

where Y_i is the log–transform of ^{222}Rn concentration at point i, $X_i^{(l)}(\boldsymbol{\Lambda}_l)$ is the l-th source predictor variable at point i with hyperparameter set $\boldsymbol{\Lambda}_l$, β_l is its source regression coefficient, and ε_i is an error term.

Variables are selected through a modified stepwise regression procedure for LUR models with multiple hyperparameter values called *A Distance Decay Regression Selection Strategy (A.D.D.R.E.S.S.)* [22]. To be more physically meaningful, all variables are considered source terms and are constrained to be positive. This model formulation supports the hypothesis that regions of elevated ^{222}Rn, or "hot spots", are due to the underlying geology and ^{238}U, and that while certain geological formations are associated with low ^{222}Rn, geological formations do not physically decrease the amount of ^{222}Rn.

BME Estimation Framework for Space/Time Mapping Analysis

To improve estimation accuracy, we integrate the time-averaged LUR results into the Bayesian Maximum Entropy (BME) method of modern spatiotemporal geostatistics [36,37]. BME is a space/time geostatistical estimation framework grounded in epistemic principles that reduces to the space/time simple, ordinary, and universal Kriging methods as its linear limiting case when considering a limited, Gaussian, knowledge base, while also allowing the flexibility to process a wide variety of additional knowledge bases (physical laws, empirical relationships, non-Gaussian distributions, hard and soft data, etc.). We only provide the fundamental BME equations for mapping ^{222}Rn. The reader is referred to other works for more detailed derivations of BME equations [36,38] and the LUR integration into BME [25].

Let $Z(\boldsymbol{p})$ be the space/time random field (S/TRF) describing the distribution of groundwater log-^{222}Rn across space and time, where $\boldsymbol{p} = (\boldsymbol{s}, t)$, $\boldsymbol{s}$ is the space coordinate and t is time. The knowledge available is organized in the general knowledge base (G-KB) about the space/time trend and variability (e.g. mean, covariance) of ^{222}Rn across the study domain, and the site-specific knowledge base (S-KB) corresponding to the hard and soft data $\boldsymbol{z_d}$ available at a set of specific space/time points $\boldsymbol{p}_d$.

First, we define the transformation of log-^{222}Rn data $\boldsymbol{z_d}$at locations $\boldsymbol{p_d}$ as

$$\boldsymbol{x_h} = \boldsymbol{z_h} - o_Z(\boldsymbol{p_h}) \tag{2.21}$$

where $o_Z(\boldsymbol{p_h})$ may be any deterministic offset that can be mathematically calculated at any space/time coordinate $\boldsymbol{p}$. We then define $X(\boldsymbol{p})$ as a homogeneous/stationary S/TRF representing the variability and uncertainty with the transformed data $\boldsymbol{x_d}$, i.e. such that $\boldsymbol{x_d}$ is a realization of

$X(\boldsymbol{p})$. Finally we let $Z(\boldsymbol{p}) = X(\boldsymbol{p}) + o_z(\boldsymbol{p})$ be the S/TRF representing groundwater log-^{222}Rn. In this study, we consider two choices for $o_z(\boldsymbol{p})$**:** (1) a constant value determined by the mean resulting in a purely BME model, and (2) the LUR estimate $\boldsymbol{L_z(p_h)}$ resulting in a LUR-BME model.

The G-KB for the S/TRF $X(\boldsymbol{p})$ describes its local space/time trends and dependencies. In this work, the general knowledge consists of the space/time mean trend function $m_x(\boldsymbol{p}) = E[X(\boldsymbol{p})]$, and the covariance function $C_X(\boldsymbol{p}, \boldsymbol{p}')$=$E[[X(\boldsymbol{p}) - m_x(\boldsymbol{p})][\mathrm{X}(\mathbf{p}') - m_x(\boldsymbol{p}')]]$ of the S/TRF $X(\boldsymbol{p})$. We calculate isotropic and anisotropic experimental covariance values at four directions of azimuth (0, 45, 90, 135). Additionally, we divide the BME and LUR-BME analysis into 3 physiographic provinces (Figure 2.1) of North Carolina based on geological properties: Blue Ridge, Piedmont, and Coastal Plain. The covariance is modeled by physiographic region if there are significant differences in model parameters between each region. Furthermore, the principal anisotropic axis is determined by examination of the experimental covariance plots and the major axis of an ellipse fit to a rose diagram: a plot of the spatial experimental covariance range as a function of the azimuth. For anisotropic models, the range of the covariance is always the range of the model along the principal axis and coordinates are converted from the anisotropic to isotropic case.

S-KB consists of hard data and soft data; with hard data, $\boldsymbol{x_h} = \boldsymbol{z_h} - \boldsymbol{L_z(p_h)}$, for data points where $\boldsymbol{z_h}$ is observed over the detection limit and soft data, $\boldsymbol{X_s}$, is at locations $\boldsymbol{p_s}$where ^{222}Rn is observed below the detection limit. Following Messier et al [25,28], the BME soft data for log-^{222}Rn is modeled as a Gaussian distribution truncated above the log of the detection limit.

The overall knowledge bases considered consist of $G = \{m_x(\boldsymbol{p}), C_X(\boldsymbol{p}, \boldsymbol{p}')\}$, and $S = \{f_s(.), \boldsymbol{X_h}\}$. In this case the BME set of equations reduces to

$$f_K(x_k) = A^{-1} \int d\boldsymbol{x_s} f_G(\boldsymbol{x_h}, \boldsymbol{x_s}, x_k) f_S(\boldsymbol{x_s}) \tag{2.22}$$

where $f_K(x_k)$ is the BME posterior PDF for the offset-removed log-$^{222}Rn(x_k)$ at some unmonitored estimation point $\boldsymbol{p}_k$, $f_G(\boldsymbol{x_h}, \boldsymbol{x_s}, x_k)$ is the (maximum entropy) multivariate Gaussian PDF for $(\boldsymbol{x_h}, \boldsymbol{x_s}, x_k)$ with mean and variance-covariance given by G-KB, $f_S(\boldsymbol{x_s})$ is the truncated Gaussian PDF of $\boldsymbol{X_s}$, and $A^{-1}$is a normalization constant. After the BME analysis is conducted, $o_Z(\boldsymbol{p})$ is added back to obtain log-^{222}Rn concentrations.

Validation Statistics

Results between LUR, BME, and LUR-BME are compared with a leave-one-out cross-validation. In LOOCV, each log-^{222}Rn value Z_j is removed one at a time, and re-estimated using the given model based only on the remaining data. We assess the accuracy and precision with the Root Mean Squared Error (RMSE), the precision with R^2, and the bias of the estimated standard deviation with the Root Mean Squared Standardized Error (RMSS). Let $Z^{*(k)}$be the re- estimate for method *k*, then $RMSE^{(k)} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(Z_j^{*(k)} - Z_j\right)^2}$, the cross-validation R-Squared is $R^2(\boldsymbol{Z}, \boldsymbol{Z}^{*(k)})$**,** and the $RMSS^{(k)} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(Z_j^{*(k)} - Z_j\right)^2 / \hat{\sigma}_j^{*(k)}}$, where $\hat{\sigma}_j^{*(k)}$ is the prediction standard error. RMSS should be close to one if the prediction standard errors are valid.

Kruskal-Wallis Hypothesis Tests for LUR model results

A major goal of this study is to help elucidate intra-geological differences that result in local groundwater ^{222}Rn variability; or to explain anomalies in which a *general* geological description is generally associated with elevated ^{222}Rn, but contains an *element* or *unit* that is associated with low ^{222}Rn. The geology and uranium based explanatory variables and the geological classification scales allow us to generate and test hypotheses from our LUR model results. To this end, we perform a Kruskal-Wallis non-parametric ANOVA test [39] on the ^{238}U or ^{222}Rn concentrations within geological formations that were selected to the final LUR model. For instance, if a *general* classification scale variable is selected with a geology and uranium based variable and there is a *element* or *unit* classification scale geological formation that is a subset of the *general* variable with low observed ^{222}Rn concentrations, then we can compare the distributions of the ^{238}U concentrations within the subset geological formation to the larger group, to statistically test if the ^{238}U is significantly higher in the larger group then the subset, thereby driving the larger group's intra-geological ^{222}Rn variability. Similarly, we can compare ^{222}Rn distributions in geological formations and their subset formations if both were selected (i.e. *Element vs. General),* testing whether subset formations contribute ^{222}Rn concentrations to groundwater in the larger group to varying degrees. The Kruskal-Wallis does not make an assumption on the normality of the data, and the null hypothesis is that the two groups come from the same distribution with equal medians.

Results

Land Use Regression

The results including the geological scale, ^{238}U chemistry, ellipse size and angles, linear coefficients, and p-values of the LUR model selected by A.D.D.R.E.S.S. for groundwater ^{222}Rn are summarized in table 2.1. With 15 explanatory variables selected plus an intercept, the model obtained a R^2 of 0.33 (Pearson correlation coefficient= 0.57). The LUR maps of predicted ^{222}Rn median and variance are available in the supplemental material.

Table 2.12. Land Use Regression model selected through A Distance Decay Regression Selection Strategy.

Variable	Geological Scale	Chemistry/Percent	Ellipse (major, minor, angle)	Beta	P-Value
Intercept	-	-	-	6.0829	0
Intrusive Felsic	General	Sediment Uranium	10km/10km/-	0.0470	0.0152
Laurentian metasedimentary and volcanics	Unit	Sediment Uranium/Alkalinity	5km/2km/135	0.2661	1.42e-17
Piedmont Zone Eastern Blue Ridge	Element	Percent	10km/7.5km/180	0.0092	1.59e-20
Grandfather Mountain Window	Unit	Sediment Uranium	10km/5km/45	0.5487	6.84e-13
Carolina Zone Raleigh Terrane	Element	Percent	10km/2.5km/135	0.0207	2.79e-15
Cherryville Pluton	Unit	Percent	7.5km/2.5km/90	0.0251	0.0018
Milton Terrane	Unit	Groundwater Uranium/Conductivity	7.5km/2.5km/180	0.6666	1.37e-7
Beech Pluton	Unit	Groundwater Uranium/Conductivity	10km/1km/90	54.75	4.23e-10

Deep River Basin	Element	Sediment Uranium/ Conductivity	7.5km/2.5km/180	40.48	1.18e-9
Piedmont Zone Tugaloo	Element	Percent	7.5km/1km/135	0.0135	4.27e-11
Late Paleozoic Plutons	Element	Percent	5km/5km/-	0.0181	3.90e-29
Henderson Gneiss	Unit	Percent	10km/7.5km/135	0.0300	3.79e-25
Mecklenburg Pluton	Unit	Groundwater Uranium/Conductivity	7.5km/2.5km/90	2.814	3.67e-6
Piedmont Zone Eastern Blue Ridge Plutons	Element	Percent	5km/2.5km/90	0.0089	3.73e-6
Piedmont Zone Cat Square Terrane Plutons	Element	Percent	7.5km/2.5km/180	0.0262	5.31e-5

Spatial Covariance Analysis

The purely BME analysis, with an offset of the global log-^{222}Rn mean, was modeled using an anisotropic covariance model with an additive two exponential covariance model for the Blue Ridge and Piedmont physiographic regions and an isotropic additive two exponential covariance model for the coastal plains region. Significant differences between the sill (i.e. total variance) and covariance range were found between physiographic regions, which justifies using separate covariance models by region. Additionally, the covariance range differed significantly for the Blue Ridge and Piedmont regions. The model parameters shown below were fit with a least-squared approach:

$$C_X(r) = c_1 \exp\left(-\frac{3r}{a_{r_1}}\right) + c_2 \exp\left(-\frac{3r}{a_{r_2}}\right) \tag{2.23}$$

where the first component of the sill, c_1= 1.31 $(log - pCi/L)^2$, 1.46 $(log - pCi/L)^2$, and 1.52 $(log - pCi/L)^2$ for the blue ridge, piedmont, and coastal plain physiographic regions respectively; the first spatial covariance range, a_{r_1} = 1,170 m, 767 m, and 1113 m for the three

physiographic regions respectively; the second component of the sill, c_2 = 0.52 $(log - pCi/L)^2$, 0.70 $(log - pCi/L)^2$, and 0.089 $(log - pCi/L)^2$ for the three physiographic regions respectively; and the second spatial covariance range, a_{r_2}= 206 km, 77 km, and 2399 km respectively. The principal axes of anisotropy are 45 and 90 degrees for the Blue Ridge and piedmont physiographic regions respectively. The BME covariance model plots and rose diagrams are available in the supplemental material.

The LUR-BME residual covariance lacks anisotropy in all 3 physiographic regions (See supplemental material), likely due to the elliptical based variables in the LUR model. The model parameters for the LUR-BME residual covariance are also fit with a least-squared approach and are detailed as follows: c_1= 1.31 $(log - pCi/L)^2$, 1.37 $(log - pCi/L)^2$, and 1.46 $(log - pCi/L)^2$ for the blue ridge, piedmont, and coastal plain physiographic regions respectively; the first spatial covariance range, a_{r_1} = 881 m, 1,117 m, and 1113 m for the three physiographic regions respectively; the blue ridge physiographic is a one component exponential model; the second component of the sill, c_2 = 0.11 $(log - pCi/L)^2$ and 0.07 $(log - pCi/L)^2$ for the piedmont and coastal physiographic regions respectively; and the second spatial covariance range, a_{r_2}= 14.96 km and 14.98 km respectively.

Land Use Regression – Bayesian Maximum Entropy

The LUR model was integrated as the global offset to create a LUR-BME model, which resulted in a LOOCV R^2 of 0.46 (Pearson correlation coefficient= 0.68), a 28 percent improvement over LUR, and a 4 percent improvement over BME, which obtained a R^2 of 0.44 (correlation=0.66). Figure 2.3 maps the point-level groundwater ^{222}Rn median concentration and variance across North Carolina. The cross-validation results for the LUR, BME, and LUR-BME models are summarized in table 2.2.

Table 2.13. Leave-One-Out Cross-Validation statistics for the LUR, BME, and LUR-BME methods for estimation of point-level log-$\boldsymbol{^{222}Rn}$. Units for RMSE = (log-pCi/L); $\boldsymbol{R^2}$, RMSS = unitless.

Method	RMSE	R^2	RMSS
LUR	1.20	0.33	0.82
BME	1.01	0.44	1.22
LUR-BME	0.99	0.46	1.20

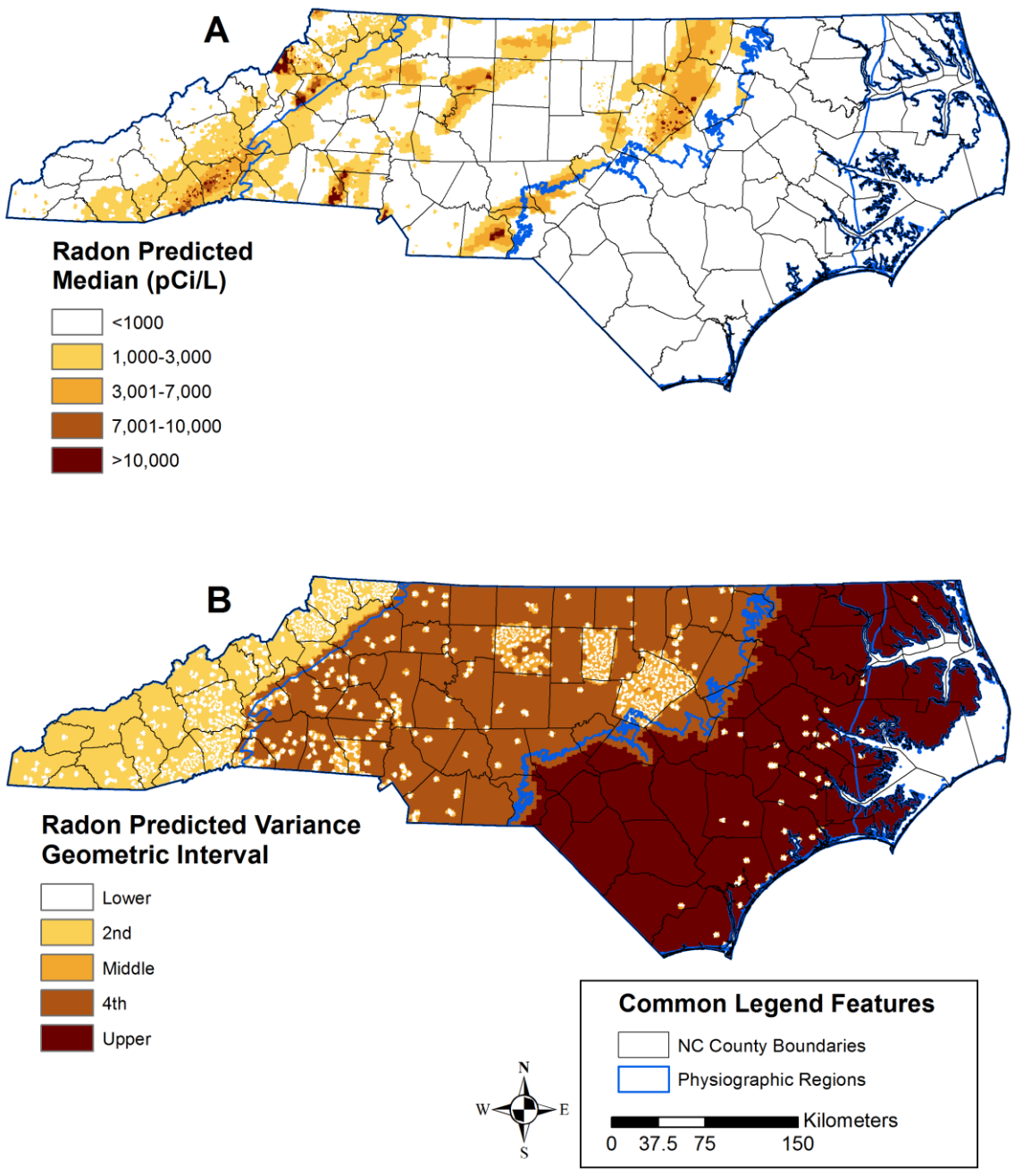


Figure 2.17. A) LUR-BME radon predicted median across North Carolina. B) LUR-BME predicted variance binned according to 5 geometric intervals. Geometric intervals are roughly quintiles, but produce better visualization for non-normal distributions.

Kruskal-Wallis ANOVA

The first variable selected in the final LUR model was the mean sediment ^{238}U within the Intrusive Felsic *general* geological formations, which contains many geological *units* known to have elevated groundwater ^{222}Rn. However, the *Greensboro Intrusive Suite* is an Intrusive Felsic *unit* that has low groundwater ^{222}Rn levels. In order to explore why the Greensboro Intrusive Suite unit has different ^{222}Rn levels then its parent Intrusive Felsic formation, we performed a Kruskal-Wallis ANOVA test on the distributions of sediment ^{238}U within the Greensboro Intrusive Suite versus the rest of the Intrusive Felsic geology. The null hypothesis is rejected with a p-value of 0, demonstrating significant difference in the distribution of uranium ^{238}U between the Greensboro Intrusive Suite and other Intrusive Felsic geologies.

The *unit* scale Henderson Gneiss, also classified as Intrusive Felsic, was selected to the final LUR model as a percent geology variable. A Kruskal-Wallis ANOVA test of observed ^{222}Rn distributions within Henderson Gneiss versus other Intrusive Felsic was rejected with a p-value of 1.7E-11, indicating an underlying higher distribution of ^{222}Rn within subcategories of Intrusive Felsic geology such as Henderson Gneiss.

Discussion

Groundwater Radon Maps

This study presents a LUR model for point-level ^{222}Rn concentration across North Carolina that elucidates geological and chemical processes affecting its variability, and then utilizes the strengths of BME to create the first map of point-level ^{222}Rn concentrations and its prediction uncertainty. Several major findings can be deduced from the first point-level groundwater ^{222}Rn maps of concentration and uncertainty across North Carolina: First, several areas of high susceptibility to elevated ^{222}Rn as determined by others [8,40] are confirmed, including the areas underlain by Henderson Gneiss (Henderson County; Blue Ridge physiographic province) and Rolesville Batholith (Eastern Wake County; Piedmont physiographic province). Second, the uncertainty is the highest in the Coastal Plain physiographic province due to the lack of data; however, there is no area with a predicted median above 3,000 pCi/L. While certainly useful, monitoring groundwater ^{222}Rn in the Coastal Plains physiographic province of North Carolina is not a high priority given the scarce state resources. Third, it would be prudent to allocate some of these scarce resources for increased monitoring in the Piedmont physiographic province in areas underlain by the Deep River basin *element* and

Late Paleozoic plutons (*element)* (see the *Element* map in supplementary materials). Our map is the first to predict (Figure 2.3A) elevated ^{222}Rn above 3,000 pCi/L almost ubiquitously across the Deep River basin and some areas above 10,000 pCi/L in Anson County due to its inclusion as an explanatory variable. However, these predictions have high uncertainty (Figure 2.3B) and are underlain with explanatory variables that greatly exceeded values used in the calibration of the model. For instance, the maximum value of the Deep River basin variable (Conductivity normalized sediment ^{238}U*)* used in calibration was 0.05 $ppb/(\mu\Omega/(cm\ ))$ whereas the maximum value found in the extrapolation of the LUR model was 0.72 $ppb/(\mu\Omega/(cm\ ))$. We limited the maximum value of explanatory variables in extrapolated regions to the maximum of the calibration range; nonetheless, high values are predicted due to the large value of its linear regression coefficient (Table 2.1). Fourth, our map predicts new areas in the Blue Ridge physiographic province with elevated groundwater ^{222}Rn including areas underlain by the Beech pluton (*Unit)* and Grandfather Mountain Window (*Unit*). The Beech pluton is also an Intrusive Felsic formation, which is known to be associated with elevated groundwater ^{222}Rn*;* however, the likely reason both units were selected in the model was their vicinity to high monitoring values in areas outside their spatial range. The Beech pluton itself only has one monitoring value within its area; however multiple high values are directly north and hence the Beech pluton was selected as a long, thin ellipse with a 90 degree azimuth.

LUR Model Interpretations

Our LUR-BME model was the first geostatistical model to account for geometric anisotropy of a groundwater contaminant through a LUR model. Our LUR model can be thought of as a groundwater version of Saito and Goovaerts [41] Kriging model for cadmium in air using predominant wind direction as a LUR variable that accounts for geometric anisotropy.

The LUR model not only sheds light on important variables in the control of groundwater ^{222}Rn, but it also allows comparison and distinction between scales of geological formations. For instance, we found the *general* geological formations of Intrusive Felsic to be important via its inclusion in the final model; moreover, Henderson Gneiss, Cherryville pluton, and Beech pluton are more detailed geologic *units* that are also Intrusive Felsic and included in the final model. The linear, additive formulation of the LUR model allowed Intrusive Felsic to be included and provide a baseline for elevated ^{222}Rn levels across much of North Carolina, which is then refined by *units* with varying local effects based on their coefficient values. Additionally, the *element*

scale variable Late Paleozoic plutons and Piedmont Zone Eastern Blue Ridge plutons were selected, which are also at least partly Intrusive Felsic. Lastly, the difference in scales allows for more significant extrapolation of potential elevated ^{222}Rn areas. As previously mentioned, Late Paleozoic plutons was selected, which contains the area of elevated ^{222}Rn in Eastern Wake County known as the Rolesville Batholith. If Rolesville Batholith was selected instead of Late Paleozoic plutons, then there would be less extrapolated high values; but given the selection of Late Paleozoic plutons, areas in Anson County and Northwestern Guilford County also have higher predicted values. This information can provide useful guidance on prioritizing areas for new monitoring.

Hypothesized Controls of Radon Anomalies

Our LUR model results help guide appropriate hypothesis tests to conduct about potential controls of radon anomalies. For instance, Campbell et al.[8] and Vinson et al. [40] both noted positive associated between elevated ^{222}Rn and Intrusive Felsic formations; however, Campbell et al. notes the apparent anomaly of the Greensboro Intrusive Suite, which has low levels of groundwater ^{222}Rn despite being Intrusive Felsic. The Kruskal-Wallis ANOVA between sediment ^{238}U was rejected, which means there is significant difference between the distributions of sediment ^{238}U within Intrusive Felsic formations with elevated ^{222}Rn and Intrusive Felsic formations with low ^{222}Rn, leading us to hypothesize that intra-geological variability of groundwater radon in Intrusive Felsic formations is at least partially controlled by sediment ^{238}U concentrations. This hypothesis is supported by Vinson et al. [40] from data in the Rolesville Batholith.

Recommendations and Limitations

This study presents a novel method whose result is a point-level mapping with physical interpretations. Human health related recommendations based on the results should however consider the limitations of the study. The results of this study can be used as the exposure assessment in a retrospective epidemiological analysis as this represents the best estimate currently available for groundwater ^{222}Rn concentrations in North Carolina; but, there is the potential for exposure misclassification, especially in areas outside the spatial covariance range. However, LUR-BME also provides the benefit of an accurate quantification of uncertainty (RMSS=1.20) to use is a risk assessment framework.

Groundwater ^{222}Rn was observed at the point-level, the theoretical lower limit of the scale of our LUR-BME estimates; however, the geological information used in the study [29] was at the 1:1500,000 map scale, which results in a theoretical lower limit for detectable size of 1500 meters and a raster grid cell size of 750 meters [42]. This along with the paucity of data limits drawing conclusions on the effects of geology at the local scale (10^1 m). The LUR-BME method presented in this paper is however easily translatable to smaller areas. For instance, the USGS creates geological "Quadrangle" maps at the 1:24,000 map scale, which would allow scales larger than the average parcel to be resolved given sufficient monitoring data as well. We considered using the 1:24,000 quad maps for this study; however they do not cover the entire study domain, and the level of detail is too refined and results in a majority of zeroes or null explanatory variables. Nonetheless, given sufficient groundwater ^{222}Rn samples in a local area, the quad maps could be used with our method to model point-level variability and elucidate local scale effects of geology.

Groundwater ^{222}Rn has been shown to be positively correlated ($R^2 = 0.37$) with well and casing depth [40], but this information was only available for a small subset of our data (< 10%). Neither casing nor well depth were considered as potential explanatory variables; however, given that the geological information is also two dimensional, depth information would not elucidate additional geological controls.

Conclusions

A LUR model with novel anisotropic explanatory variables can elucidate important predictors of point-level groundwater ^{222}Rn in North Carolina. The methods are translatable to other study areas in the United States and to different spatial scales. LUR-BME models can be used to predict spatial varying groundwater ^{222}Rn and provide uncertainty assessments. Kruskal-Wallis ANOVA hypothesis tests help explain intra-geological differences of ^{222}Rn concentrations due to the occurrence of ^{238}U. Further research on ^{222}Rn health effects such as retrospective epidemiological analyses can use our results as the exposure assessment. Lastly, results will be useful in identifying localities of elevated ^{222}Rn for increased monitoring and areas with little to no monitoring that need to be monitored due to their predicted potential for elevated groundwater ^{222}Rn.

REFERENCES

(1) WHO. *WHO guidelines for drinking-water quality.*; 2011.

(2) Darby, S.; Hill, D.; Auvinen, A.; Barros-Dios, J. M.; Baysson, H.; Bochicchio, F.; Deo, H.; Falk, R.; Forastiere, F.; Hakama, M.; et al. Radon in homes and risk of lung cancer: collaborative analysis of individual data from 13 European case-control studies. *BMJ* **2004**, *330*, 223.

(3) Field, R. W.; Smith, B.; Steck, D.; Lynch, C. F. Residential radon exposure and lung cancer: variation in risk estimates using alternative exposure scenarios. *J. Expo. Anal. Environ. Epidemiol.* **2002**, *12*, 197–203.

(4) Kendall, G. M.; Smith, T. J. Doses to organs and tissues from radon and its decay products. *J. Radiol. Prot.* **2002**, *22*, 389–406.

(5) Lubin, J. H.; Boice, J. D. Lung cancer risk from residential radon: meta-analysis of eight epidemiologic studies. *J. Natl. Cancer Inst.* **1997**, *89*, 49–57.

(6) Field, R. W. Environmental Factors in Cancer: Radon. *Rev. Environ. Health* **2010**, *25*, 33–38.

(7) Krewski, D.; Lubin, J. H.; Zielinski, J. M.; Alavanja, M.; Catalan, V. S.; Field, R. W.; Klotz, J. B.; Letourneau, E. G.; Lynch, C. F.; Lyon, J. I.; et al. Residential Radon and Risk of Lung Cancer. *Epidemiology* **2005**, *16*, 137–145.

(8) Campbell, T.; Mort, S.; Fong, F.; Crawford-Brown, D.; Vengosh, A.; Cornell, E.; Field, W. R. *North Carolina Radon-in-Water Advisory Committee Report*; Raleigh, North Carolina, 2011.

(9) National Research Council. *Risk Assessment of Radon in Drinking Water*; Washington D.C., 1999.

(10) Vinson, D. S.; Campbell, T. R.; Vengosh, A. Radon transfer from groundwater used in showers to indoor air. *Appl. Geochemistry* **2008**, *23*, 2676–2685.

(11) Fitzgerald, B.; Hopke, P. K. Experimental Assessment of the Short- and Long-Term Effects of Rn from Domestic Shower Water on the Dose Burden Incurred in Normally Occupied Homes. **1997**, *31*, 1822–1829.

(12) Auvinen, A.; Salonen, L.; Pekkanen, J.; Pukkala, E.; Ilus, T.; Kurttio, P. Radon and other natural radionuclides in drinking water and risk of stomach cancer: a case-cohort study in Finland. *Int. J. Cancer* **2005**, *114*, 109–113.

(13) Brutsaert, W. F.; Norton, S. A.; Hess, C. T.; Williams, J. S. Geologic and Hydrologic Factors Controlling Radon-222 in Ground Water in Maine. *Groundwater* **1981**, *19*, 407–417.

(14) Yang, Q.; Smitherman, P. E.; Hess, C. T.; Culbertson, C. W.; Marvinney, R. G.; Smith, A. E.; Zheng, Y. Uranium and radon in private bedrock well water in Maine: geospatial analysis at two scales. *Environ. Sci. Technol.* **2014**, *48*, 4298–4306.

(15) Loomis, D. P. Radon-222 Concentration and Aquifer Lithology in North Carolina. *Groundw. Monit. Remediat.* **1987**, 33–39.

(16) Zhu, H. C.; Charlet, J. M.; Doremus, P. Kriging Radon Concentration of Groundwaters in Western Ardennes. *Environmetrics* **1996**, *7*, 513–523.

(17) Skeppström, K.; Olofsson, B. A prediction method for radon in groundwater using GIS and multivariate statistics. *Sci. Total Environ.* **2006**, *367*, 666–680.

(18) Salih, I. M.; Pettersson, H. B. L.; Sivertun, A.; Lund, E. Spatial correlation between radon (222-Rn) in groundwater and bedrock uranium (238-U): GIS and geostatistical analyses. *J. Spat. Hydrol.* **2002**, *2*, 1–10.

(19) USEPA. http://water.epa.gov/lawsregs/rulesregs/sdwa/index.cfm.

(20) Kenny, J. F.; Barber, N. L.; Hutson, S. S.; Linsey, K. S.; Lovelace, J. K.; Maupin, M. A. Estimated Use of Water in the United States in 2005. *USGS Circ. 1344* **2005**.

(21) Nolan, B. T.; Hitt, K. J. Vulnerability of shallow groundwater and drinking-water wells to nitrate in the United States. *Environ. Sci. Technol.* **2006**, *40*, 7834–7840.

(22) Su, J. G.; Jerrett, M.; Beckerman, B. A distance-decay variable selection strategy for land use regression modeling of ambient air pollution exposures. *Sci. Total Environ.* **2009**, *407*, 3890–3898.

(23) Nuckols, J. R.; Beane Freeman, L. E.; Lubin, J. H.; Airola, M. S.; Baris, D.; Ayotte, J. D.; Taylor, A.; Paulu, C.; Karagas, M. R.; Colt, J.; et al. Estimating Water Supply Arsenic Levels in the New England Bladder Cancer Study. *Environ. Health Perspect.* **2011**, *1002345*.

(24) Kim, D.; Miranda, M. L.; Tootoo, J.; Bradley, P.; Gelfand, A. E. Spatial Modeling for Groundwater Arsenic Levels in North Carolina. *Environ. Sci. Technol.* **2011**, *45*, 4824–4831.

(25) Messier, K. P.; Akita, Y.; Serre, M. L. Integrating address geocoding, land use regression, and spatiotemporal geostatistical estimation for groundwater tetrachloroethylene. *Environ. Sci. Technol.* **2012**, *46*, 2772–2780.

(26) Rodríguez-Lado, L.; Sun, G.; Berg, M.; Zhang, Q.; Xue, H.; Zheng, Q.; Johnson, C. A. Groundwater arsenic contamination throughout China. *Science* **2013**, *341*, 866–868.

(27) Sanders, A. P.; Messier, K. P.; Shehee, M.; Rudo, K.; Serre, M. L.; Fry, R. C. Arsenic in North Carolina: public health implications. *Environ. Int.* **2011**, *38*, 10–16.

(28) Messier, K. P.; Kane, E.; Serre, M. L. Nitrate Variability in Groundwater of North Carolina using Monitoring and Private Well Data Models. **2014**, Manuscript in Preparation.

(29) Hibbard, J.; van Stall, C.; Rankin, D.; Williams, H. *Lithotectonic Map of Appalachian Orogen: Canada-United States of America*; Geological Survey of Canada: Map 0206A; Map Scale 1:1 500 000, 2006.

(30) U.S. Geological Survey. *National Uranium Resource Evaluation (NURE) Hydrogeochemical and Stream Sediment Reconnaissance data*; U.S. Geological Survey: Denver, CO, 2004.

(31) Kotrappa, P.; Jesters, W. A. Electret ion chamber radon monitors measure dissolved 222Rn in water. *Heal. Physiscs* **1993**, *64.4*, 397–405.

(32) USEPA. Standard Method 7500-Rn. *Standard Methods for the Examination of Water and Wastewater*, 1999, *19th editi*.

(33) Survey, U. S. G. National Water Information System http://nwis.waterdata.usgs.gov/nwis (accessed Jun 1, 2010).

(34) Dall'Aglio, M. Planning and interpretation criteria in hydrogeochemical prospecting for uranium. In *Uranium Prospecting Handbook*; Bowie, S. H. U.; Davis, M.; Ostle, D., Eds.; The Institution of Mining and Metallurgy, 1972; pp. 121–134.

(35) Wenrich-Verbeek, K. J. *Geological Exploration for Uranium Utilizing Water and Stream Sediments*; 1980.

(36) Christakos, G. A Bayesian/maximum-entropy view to the spatial estimation problem. *Math. Geol.* **1990**, *22*, 763–777.

(37) Serre, M. L.; Christakos, G. Modern geostatistics: computational BME analysis in the light of uncertain physical knowledge - the Equus Beds study. *Stoch. Environ. Res. Risk Assess.* **1999**, *13*, 1–26.

(38) Christakos, G.; Bogaert, P.; Serre, M. L. *Temporal GIS: Advanced Function for Field-Based Applications*; Springer: New York, NY, 2002.

(39) Kruskal, W. H.; Wallis, W. A. Use of Ranks in One-Criterioon Variance Analysis. *J. Am. Stat. Assoc.* **1952**, *47*, 583–621.

(40) Vinson, D. S.; Vengosh, A.; Hirschfeld, D.; Dwyer, G. S. Relationships between radium and radon occurrence and hydrochemistry in fresh groundwater from fractured crystalline rocks, North Carolina (USA). *Chem. Geol.* **2009**, *260*, 159–171.

(41) Saito, H.; Goovaerts, P. Accounting for source location and transport direction into geostatistical prediction of contaminants. *Environ. Sci. Technol.* **2001**, *35*, 4823–4829.

(42) Tobler, W. Measuring Spatial Resolution. In *Land Resources Information Systems Conference*; Beijng, China, 1988; pp. 12–16.

Supporting Information for Chapter 2

Estimation of Groundwater Radon in North Carolina using Land Use Regression and Bayesian Maximum Entropy

*Kyle P. Messier[†], Ted Campbell[‡], Phil Bradley[§], Marc L. Serre[†]**

Authors' Affiliation:

†Department of Environmental Science and Engineering, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC 27599, United States

‡ North Carolina Department of Environment and Natural Resources, Division of Water Resources, 2090 U.S. 70 Highway, Swannanoa, NC 28778, United States

§ North Carolina Geological Survey, 1620 Mail Service Center, Raleigh, North Carolina 27699, United States

***Corresponding Author:**

Marc L. Serre

Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, University of North Carolina, 1303 Michael Hooker Research Center, Chapel Hill, NC 27599

Phone: (919) 966-7014 Fax: (919) 966-7911

This supporting information contains 14 pages and 14 figures.

Maps of Hibbard Geology Data by Geological Scale

Figures S2.1-S2.3 are maps of the Hibbard [1] geological data within North Carolina and classified into three different geological scales. The maps are intended to show the differences in scale and a perspective of the data used. For detailed information on the geological data itself, we refer the reader to the referred publication:

(1) Hibbard, J.; van Stall, C.; Rankin, D.; Williams, H. *Lithotectonic Map of Appalachian Orogen: Canada-United States of America*; Geological Survey of Canada: Map 0206A; Map Scale 1:1 500 000, 2006.

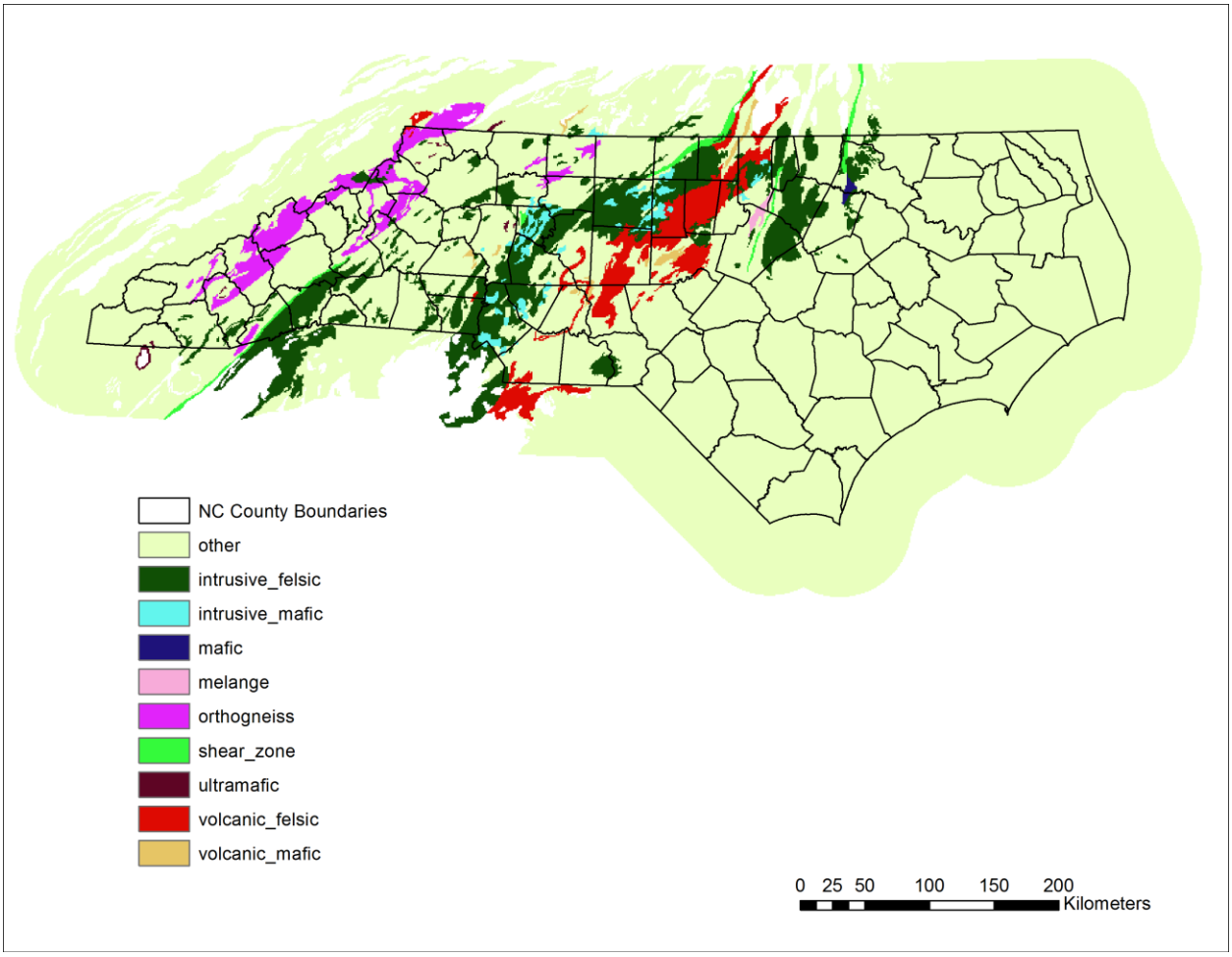


Figure S2.18. Hibbard 2006 geology data for North Carolina and surrounding 50 kilometers classified into *general* geological descriptions.

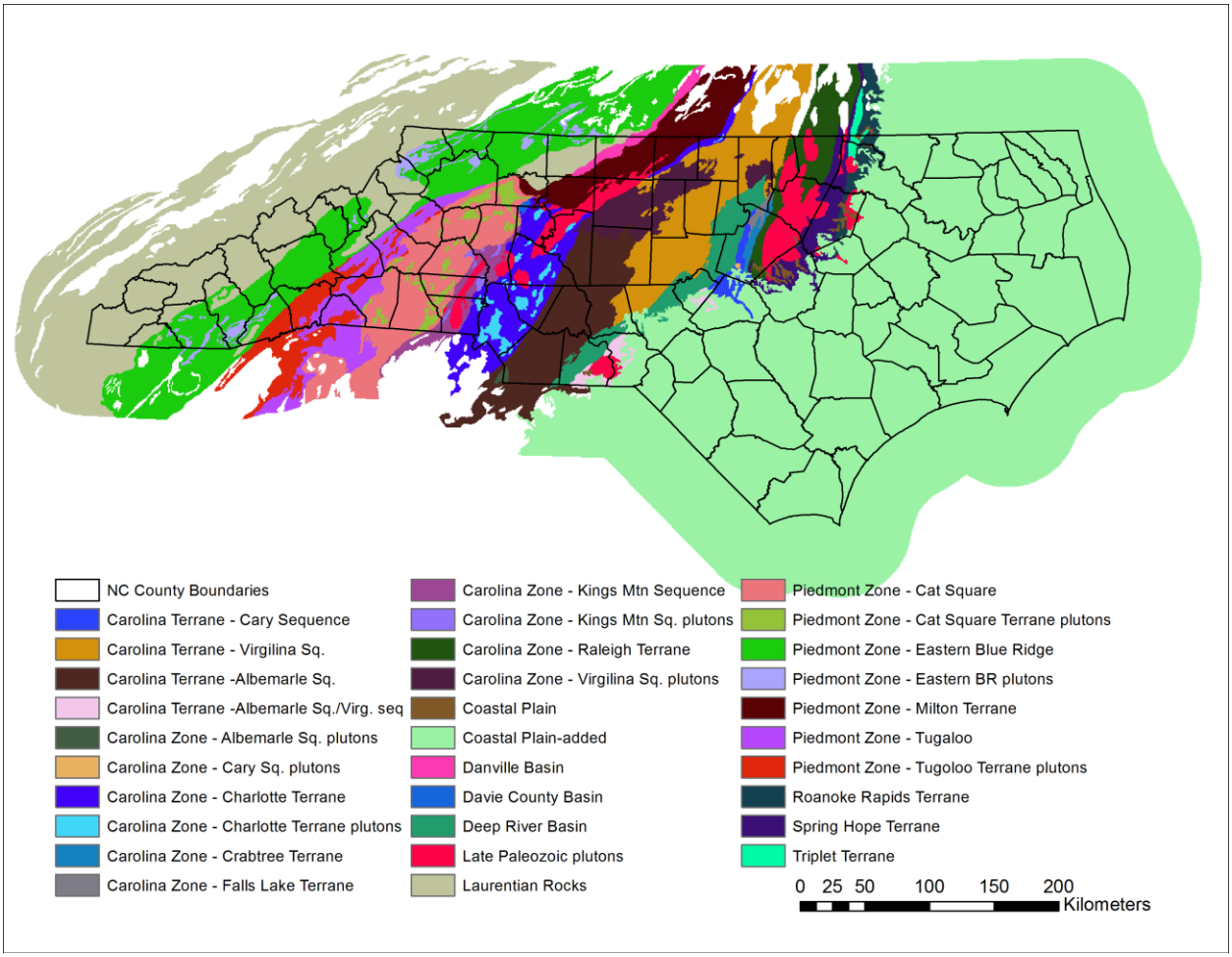


Figure S2.19. Hibbard 2006 geology data for North Carolina and surrounding 50 kilometers classified into an intermediate geological scale called *Lithotectonic elements*.

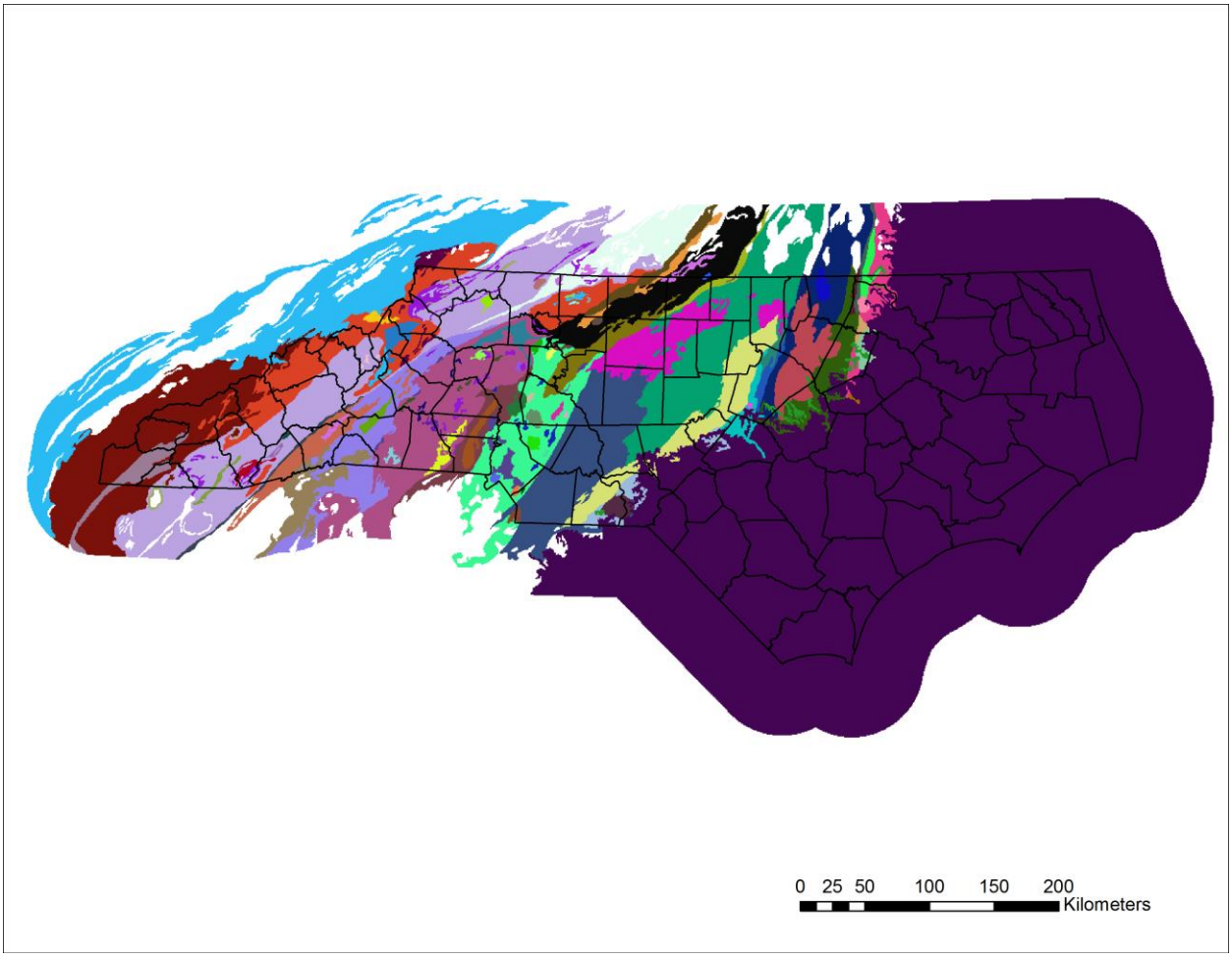


Figure S2.20A. Hibbard 2006 geology data for North Carolina and surrounding 50 kilometers classified into specific geological descriptions called *Units*.

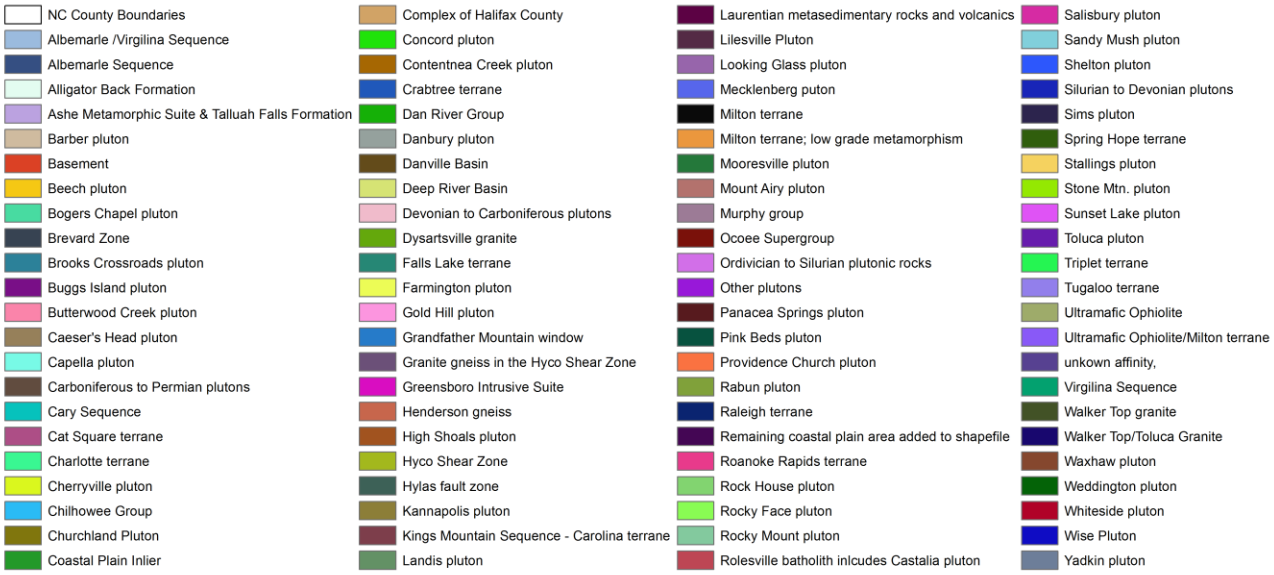


Figure S2.3B. Legend of geological *Unit* names for figure S2.3A.

Land Use Regression (LUR) Maps

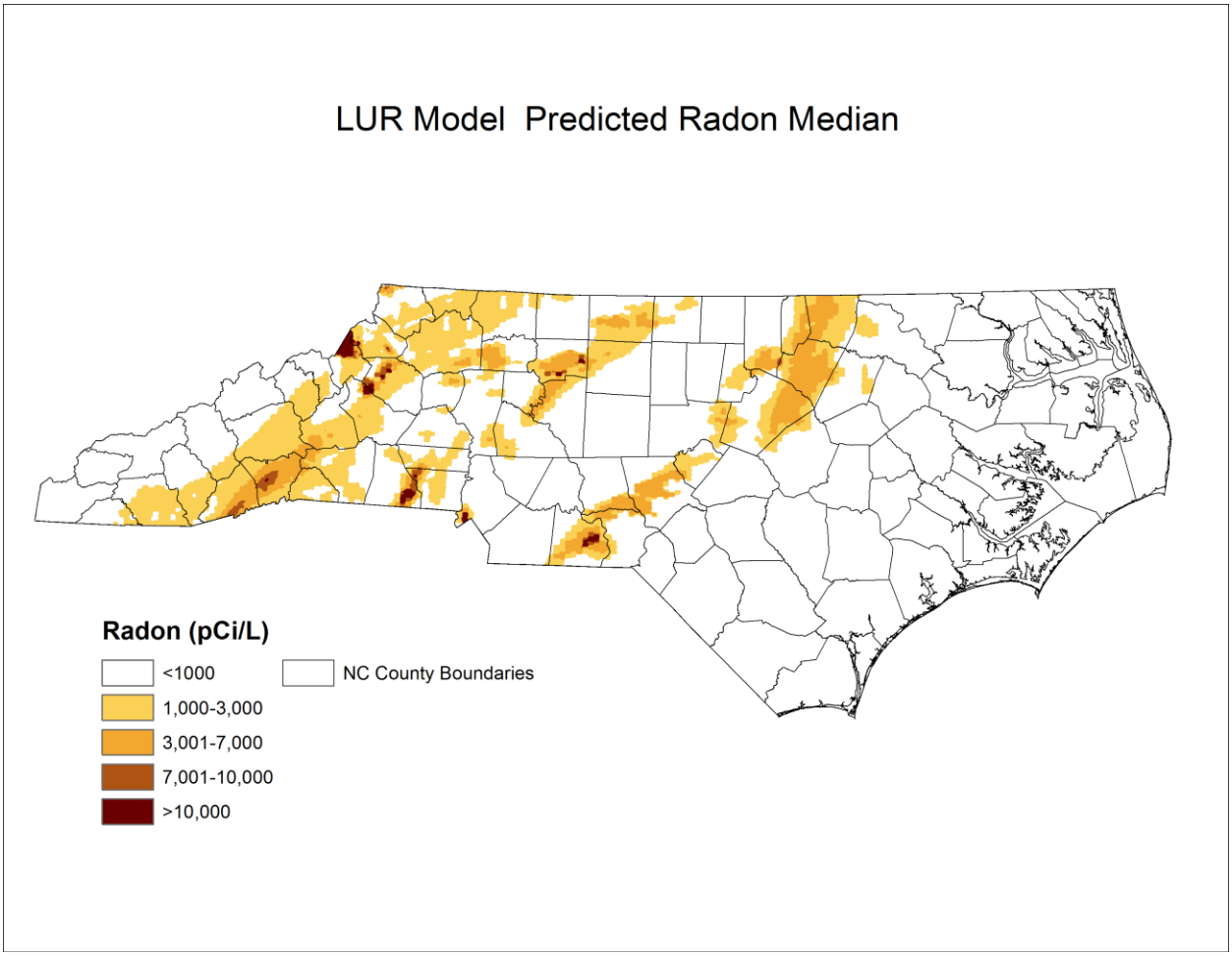


Figure S2.4. The Land Use Regression (LUR) model predicted radon median. The LUR model was selected via the A.D.D.R.E.S.S. model selection procedure.

Bayesian Maximum Entropy (BME) covariance by physiographic region

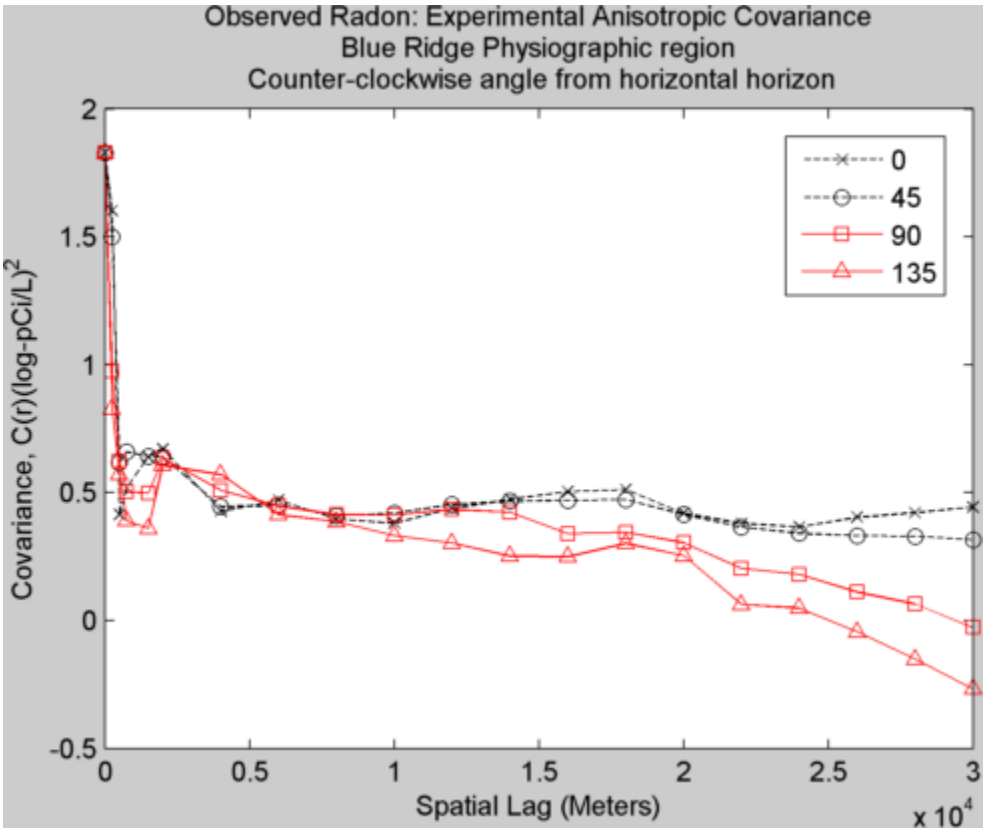


Figure S2.5. Observed radon experimental anisotropic covariance for the Blue Ridge physiographic region. Numbers represent the counter-clockwise angle from the horizontal horizon for the principal axis. There is clear anisotropy, especially starting at the 10 km spatial lag.

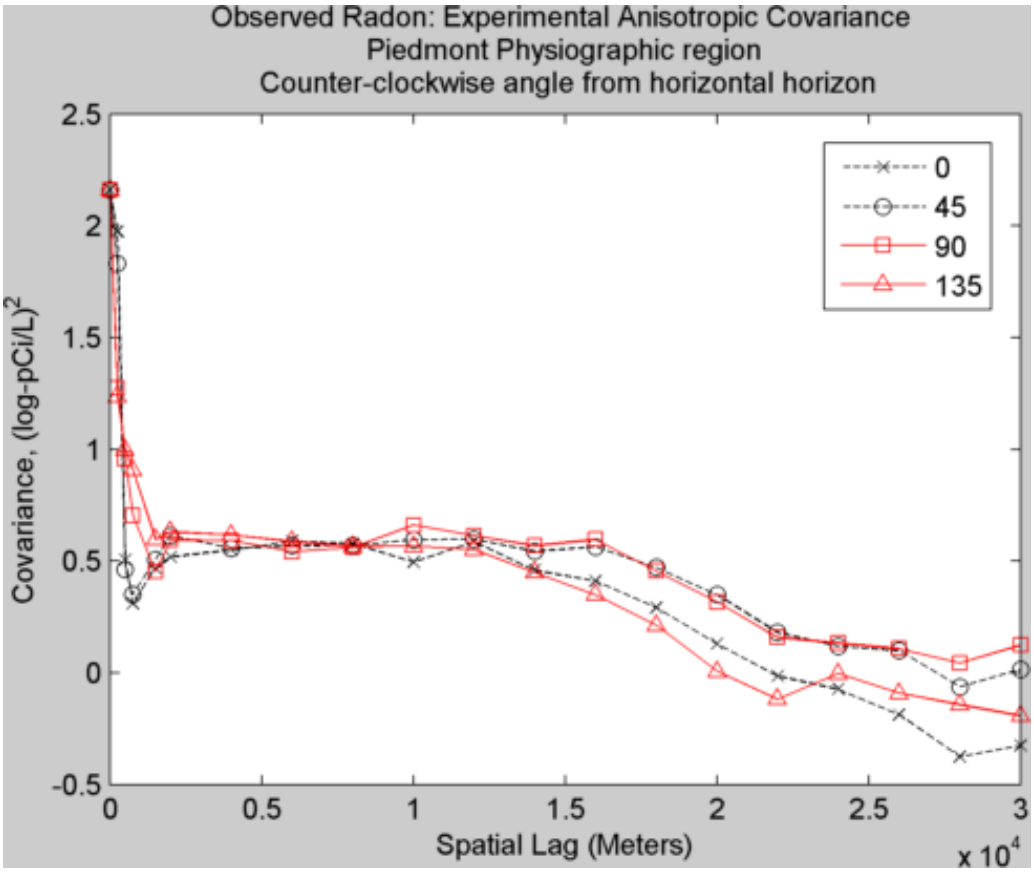


Figure S2.6. Observed radon experimental anisotropic covariance for the Piedmont physiographic region. Numbers represent the counter-clockwise angle from the horizontal horizon for the principal axis. There is clear anisotropy, especially starting at the 10 km spatial lag.

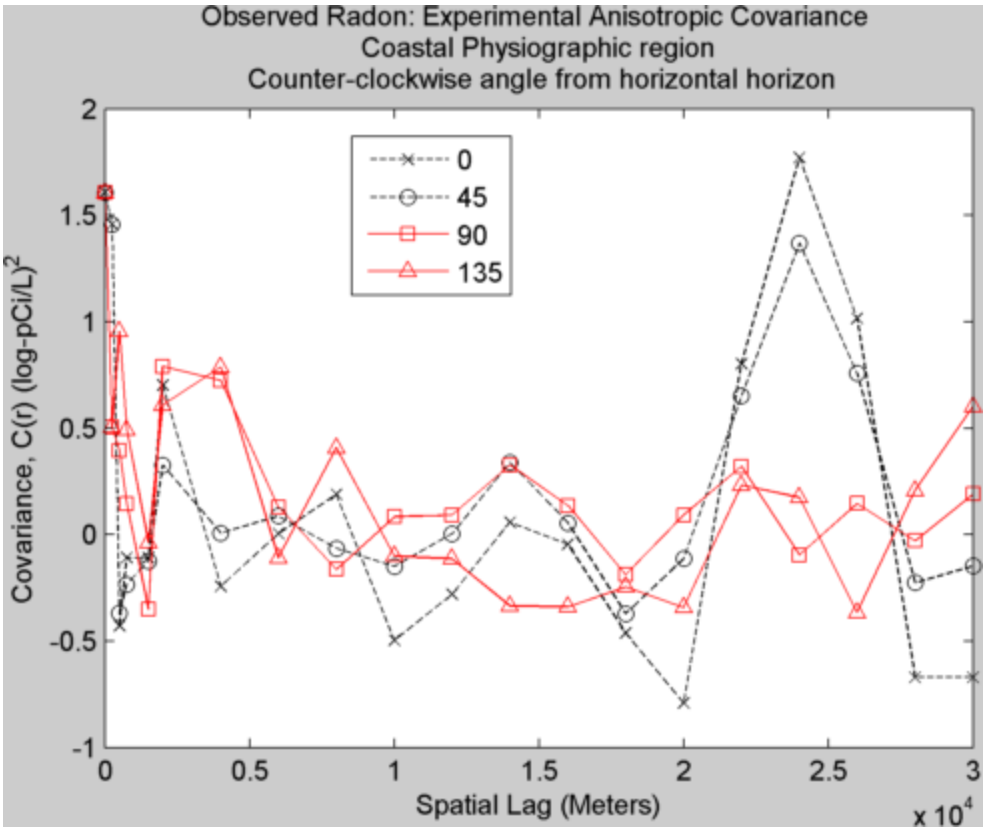


Figure S2.7. Observed radon experimental anisotropic covariance for the coastal plains physiographic region. Numbers represent the counter-clockwise angle from the horizontal horizon for the principal axis. There is no apparent difference in angle of anisotropy.

BME rose diagrams

Rose diagrams are created for both the observed radon and the LUR residual radon data. We show the diagrams for only the Blue Ridge and Piedmont physiographic regions because the coastal plains physiographic region has noisy results for experimental anisotropic covariance. Rose diagrams are created by plotting the experimental covariance value at a particular spatial lag for every given azimuth tested. Then when the results are plotted in polar coordinates and a circle or ellipse is fitted to the data. If an ellipse is fit to the data, then it indicates a major axis of geometric anisotropy. We chose to use the 16km spatial lag as the lag for the rose diagram calculations. Results are generally consistent regardless of the spatial lag chosen; however, some differences do occur due to variability in the data. Nonetheless, 16km spatial lag provides stable results.

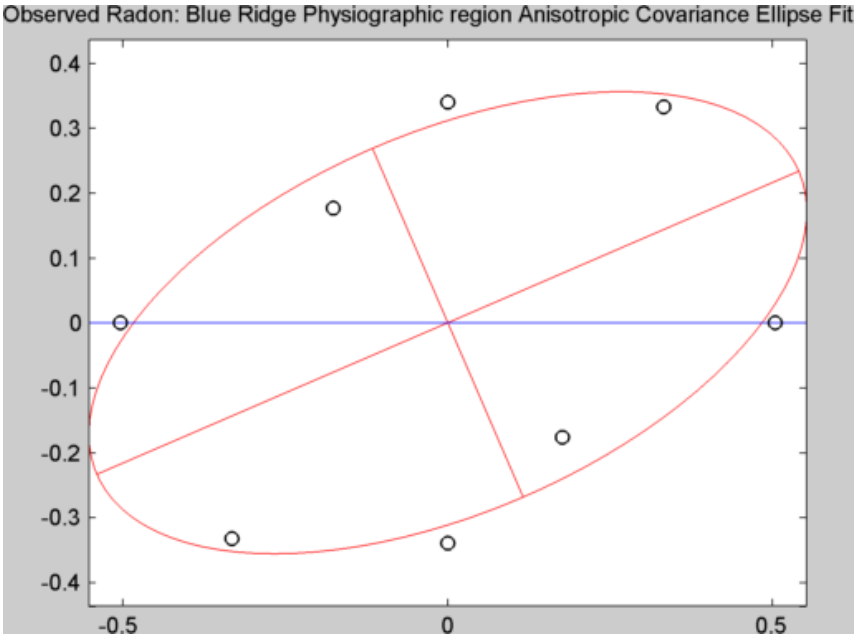


Figure S2.8. A rose diagram for observed radon within the Blue Ridge physiographic region. The major axis of the ellipse is close to the 45 degree azimuth.

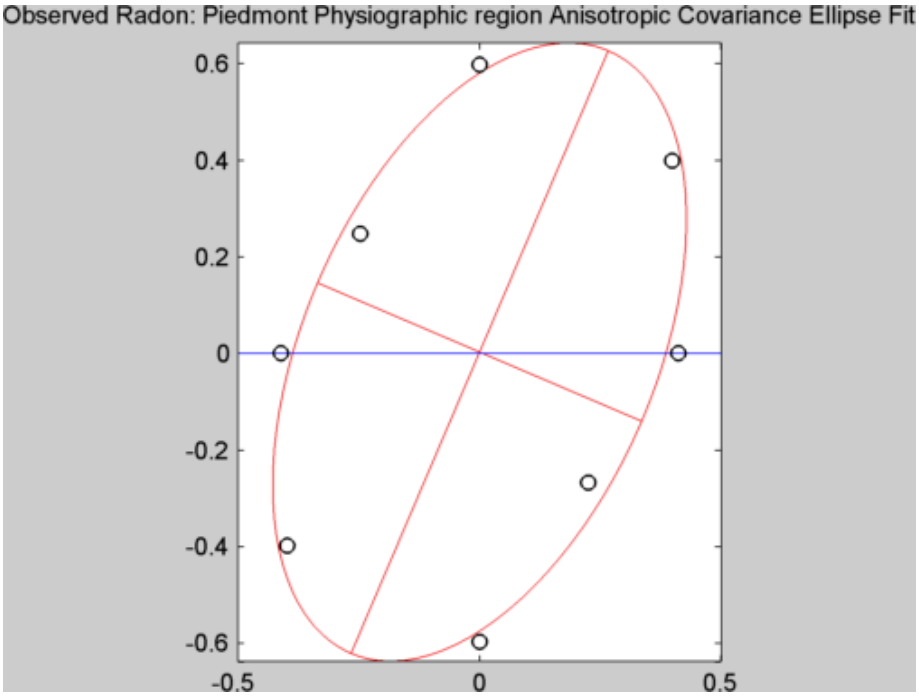


Figure S9. A rose diagram for observed radon within the Piedmont physiographic region. The major axis is close to the 90 degree azimuth.

LUR-BME covariance by physiographic region

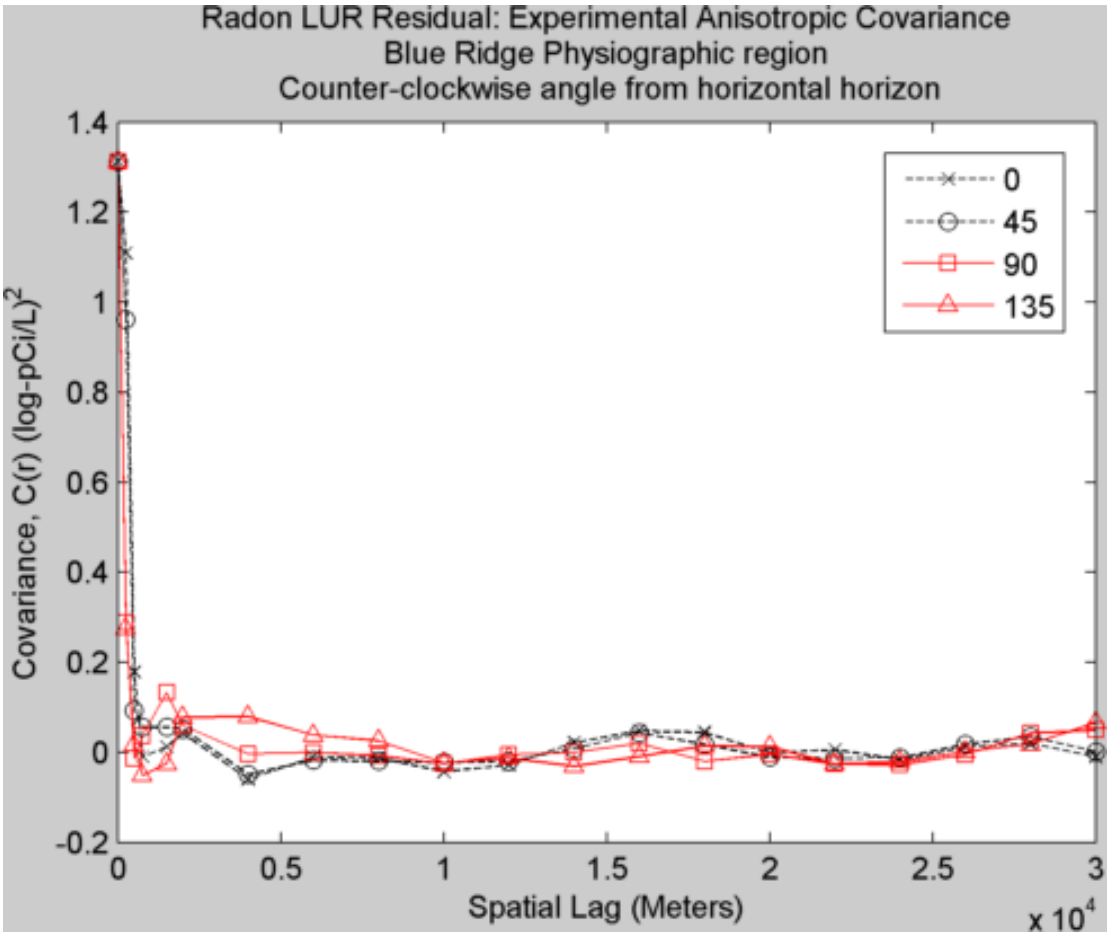
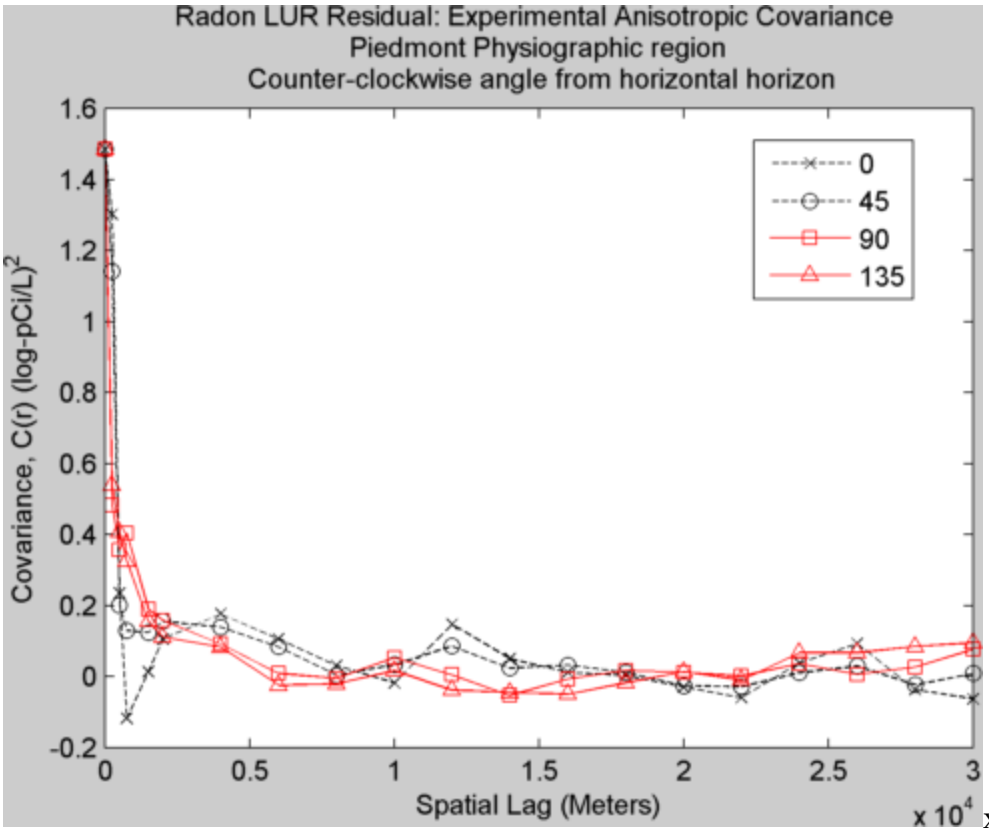


Figure S2.210. Radon LUR residual experimental anisotropic covariance for the Blue Ridge physiographic region. Numbers represent the counter-clockwise angle from the horizontal horizon for the principal a



xis. There is no apparent difference in angle of anisotropy.

Figure S2.11. Radon LUR residual experimental anisotropic covariance for the Piedmont physiographic region. Numbers represent the counter-clockwise angle from the horizontal horizon for the principal axis. There is no apparent difference in angle of anisotropy.

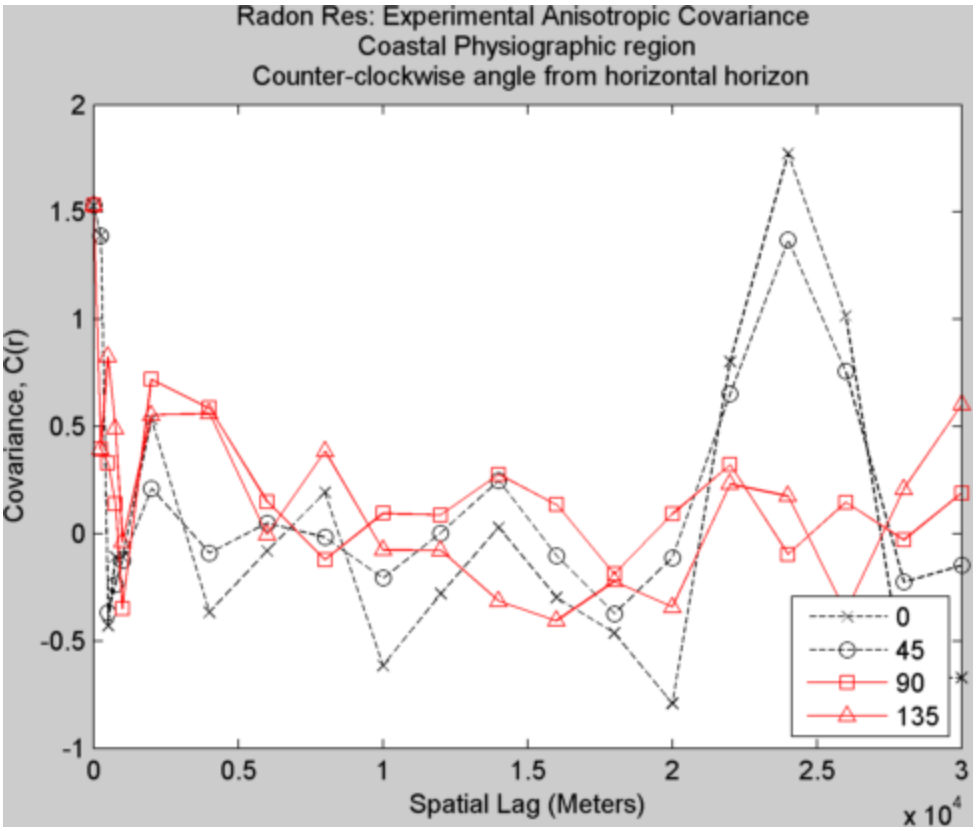


Figure S2.12. Radon LUR residual experimental anisotropic covariance for the coastal plains physiographic region. Numbers represent the counter-clockwise angle from the horizontal horizon for the principal axis. There is no apparent difference in angle of anisotropy.

LUR-BME rose diagrams

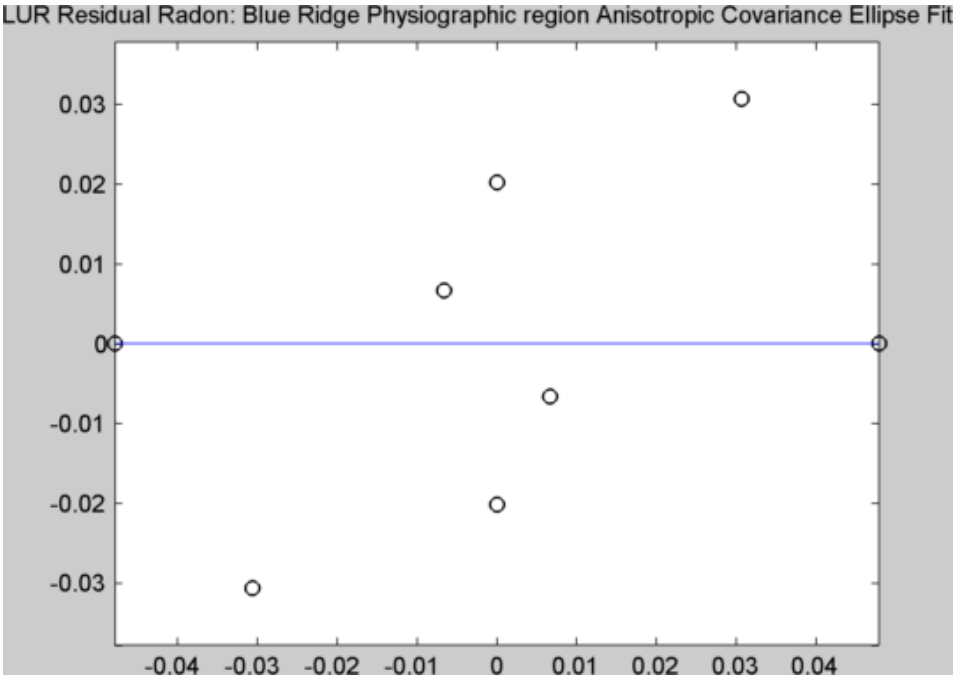


Figure S2.22. A rose diagram for radon LUR residual within the Blue Ridge physiographic region. An ellipse is not able to be fit to the data because of the lack of anisotropy.

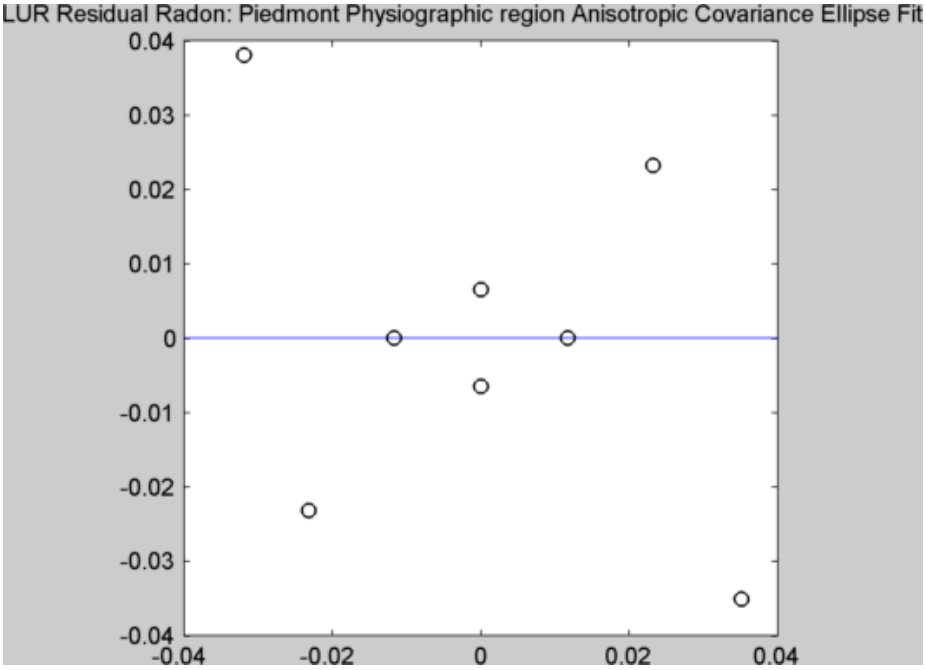


Figure S2.14. A rose diagram for radon LUR residual within the Piedmont physiographic region. An ellipse is not able to be fit to the data because of the lack of anisotropy

CHAPTER 3

Lung and Stomach Cancer Associations with Groundwater Radon in North Carolina, United States at Multiple Spatial Scales

Kyle P. Messier[†] *and Marc L. Serre*[†]*

Authors' Affiliation:

†Department of Environmental Science and Engineering, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC 27599, United States

***Corresponding Author:**

Marc L. Serre

Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, University of North Carolina, 1303 Michael Hooker Research Center, Chapel Hill, NC 27599

Phone: (919) 966-7014 Fax: (919) 966-7911

Acknowledgements:

This research was supported in part by funds from the NIH T32ES007018, NIOSH 2T42OH008673. We thank the North Carolina Central Cancer Registry and the State Center for Health statistics including Dr. Chandrika Rao, Dr. Luis Carrasco, and Christian Klaus for providing, geocoding, and geomasking the cancer data.

Abstract

Background: The risk of indoor air radon on lung cancer is well studied, but the risks of groundwater radon on both lung and stomach cancer are much less studied and with mixed results.

Methods: Geomasked and geocoded stomach and lung cancer cases in North Carolina from 1999-2009 were obtained from the North Carolina Central Cancer Registry. Models for the association with groundwater radon and multiple confounders were implemented at two scales: 1) An ecological model of cancer incidence rates at the census-tract level, and 2) An individual-level model estimating the odds that cancer cases belong to cancer clusters, consisting of a cluster analysis followed by logistic regression of case cluster membership .

Results: At the ecological-level, we find groundwater radon to be a significant and positive risk factor for lung cancer (Incidence Rate Ratio =1.05, 95% CI=1.01-1.08, for a 1 log-pCi/L increase in census tract log-concentration), and positive but insignificant risk for stomach cancer (IRR =1.02, 95% CI=(0.97,1.08)). At the address level we find that groundwater radon exposure significantly increases the odds that cancer cases are members of cancer clusters for lung cancer (OR=1.32, 95% CI=1.28-1.36) and stomach cancer (OR=1.18, 95% CI=1.07-1.31) after controlling for confounding factors.

Conclusion: Our study is the first epidemiological analysis finding a significant positive association between groundwater radon exposure and lung cancer incidence rates, and the first to find that groundwater radon increases the odds that both lung and stomach cancer cases are geographically clustered. The results corroborate previous biokinetic and mortality studies that groundwater radon is a significant environmental risk factor for lung and stomach cancer.

Keywords: Radon, Groundwater, Stomach Cancer, Lung Cancer, Generalized Linear Models, Cluster Analysis

Key Messages:

- The first epidemiology study of groundwater radon and lung cancer incidence
- The first epidemiology study of groundwater radon and stomach cancer to find a positive and significant risk
- Groundwater radon concentration is a significant risk factor associated with lung cancer incidence at the ecological and individual scale
- Groundwater radon concentration is a significant risk factor associated with stomach cancer at the individual level

Introduction

Radon is a naturally occurring radioactive gas and human carcinogen found in the groundwater drinking supply and indoor air across the world. Countries with documented groundwater radon occurrence include The United States of America[1–3], Finland[4], Belgium[5], Italy[6], and many other European countries[7]. The carcinogenic risk associated with radon exposure is due to its radioactive decay and emission of high energy alpha decay particles (α-decay)[8,9], thus when referring to Radon, it is generally understood to be Radon and its associated α-decay.

There is vast literature including multiple epidemiological analyses supporting the conclusion that exposures via inhalation of radon in indoor air lead to a significant increased risk of lung cancer morbidity in both never-smokers and smokers [7,10–14]. Ingestion of radon is also thought to be associated with lung cancer; however, the literature for groundwater or drinking-water route of exposure and lung cancer is limited to biokinetic models[8,15] and one ecological epidemiology analysis of mortality[16].

Stomach cancer is likely to be the second major cancer risk from radon exposure after lung cancer [8,9,11]; however, no study to date has both effectively and directly quantified this risk[11]. Previous studies have looked at stomach cancer and radon with mixed results. A case-cohort study of private well radon found a protective effect that was not statistically significant; however, it most likely suffered from a small cohort (n=371) and lack of confounders controlling for unmeasured protective effects [4]. A county scale ecological analysis found a positive relationship between indoor air radon and stomach cancer mortality, however the study did not report the number of subjects or the confidence intervals[17]. Kendall and Smith[11] hypothesized that the mixed results of stomach cancer studies is purely because there has not been a study with a highly exposed cohort of sufficient sample size.

North Carolina contains geological features commonly associated with elevated radon and has many areas across the state with high concentration of radon in the groundwater[3]. Furthermore, state-wide lung cancer incidence rates are higher than the national average for 2007-2011 (72.7 vs. 64.9 per 100,000 people) and near the national average for stomach cancer (6.7 vs 6.3 per 100,000 people)[18].

The objectives of our study are to: 1) Provide the first epidemiological analysis of groundwater radon exposure and lung cancer incidence and 2) Conduct the first epidemiological

analysis of groundwater radon and stomach cancer incidence with a large and exposed cohort. To this end, we develop two types of models for lung and stomach cancer in North Carolina across an eleven year period. The first type of model examines associations at an ecological scale, investigating the association of groundwater radon exposure and lung and stomach cancer incidence rates by census tract. To expand upon the ecological-level model, we develop a two-stage cluster analysis and logistic regression framework that estimates the odds that cancer cases belong to cancer clusters, which allows for an assessment at the individual as opposed to ecological scale. This framework has been applied to evaluating the associations between H5N1 avian bird flu and environmental factors[19,20], Amyotrophic lateral sclerosis and lake water quality[21], and tuberculosis and aboriginal ancestry[22].

Results will be of interest to cancer researchers across disciplines including toxicologist and epidemiologists, federal and state agencies monitoring public health such as the department of health and human services, and to the general public in order to become better educated on their potential risks associated with groundwater radon exposure. Furthermore, the results will provide the relative risk estimate needed to calculate the sample size for a large case-control study of radon and cancer outcomes, which will be significantly more expensive and time-consuming than this study.

Methods

Study Population

Geomasked address level stomach and lung incident cancer cases in North Carolina from 1999-2009 were obtained from the North Carolina Central Cancer Registry (NCCCR) with a data use agreement. An Internal Review Board (IRB) assessment was obtained (UNC-IRB #12-1761) for human subjects; however the only identifiable information is their location. Geomasked locations are moved slightly from true addresses using a donut geomask to protect privacy while preserving the sensitivity and specificity of detecting disease clusters[23,24] Attributes include race, age at diagnosis, gender (Table 1), and various notes including tobacco use history; however, those are reported in less than 10% of cases. Stages of cancer were also not included.

Table 3.14. Basic information for the study population. Lung and stomach cancer cases from 1999-2009 in North Carolina, United States.

	Stomach Cancer	Lung Cancer
Male		
White		
< 65	814	10 080
≥ 65	1 345	20 065
Black		
< 65	423	3 099
≥ 65	457	3 244
Other		
< 65	55	217
≥65	34	219
Female		
White		
< 65	413	7 663
≥ 65	960	15 083
Black		
< 65	236	1 776
≥ 65	401	2 006
Other		
< 65	41	161
≥ 65	39	191
Total	5 218	63 804

Exposure Data

Groundwater radon concentration ($\log(pCi/L)$) exposure is estimated from Messier et al. [3], which are address-level estimates of groundwater radon concentration based on the land use regression and Bayesian Maximum Entropy (LUR-BME) geostatistical model.

Statistical Analyses at Multiple Spatial Scales

Associations between stomach and lung cancer are examined at two different spatial scales:

First, incidence rates are examined at the census tract level using a negative binomial generalized linear model (GLM) with standard NB2 parameterization[25,26]. The NB2 model is a negative binomial regression model based on the Poisson-gamma mixture probability distribution function. The benefit of this parameterization is that it allows us to model Poisson heterogeneity, or more specifically in most cases, Poisson overdispersion due to excess zero counts [25]. The model of stomach or lung cancer counts, y, is assumed to follow a negative binomial distribution such that $y \sim NB2(\mu, \alpha)$, where μ is the mean, and α is the negative binomial dispersion parameter. For the NB2 parameterization the natural log is the link function and the exponential is the inverse-link, thus we model cancer counts as

$$\ln(Y) = \beta_0 + \beta_1 Z_1 + \cdots + \beta_n Z_n + \varepsilon + offset \quad (3.1)$$

where Y is the number of stomach or lung cancer counts in a given census tract over the 11 years study period, β_n are linear coefficients for the census tract predictor variables Z_n, ε is the error term, and *offset* is the population-year? offset, which is the natural log of the census tract population times the duration of the study period (11 years) with a coefficient constrained to 1 resulting in an incidence rate interpretation of the model.

The predictor variables include the exposure Z_1 of interest (the census tract average of groundwater radon log-concentration, log-pCi/L), and known confounding variables, Z_l, l>1, which include indoor air radon exposure, smoking prevalence, public water supply status, residential tenure, age, gender, and race. Indoor air radon is considered by including the United States Environmental Protection Agency (USEPA) estimates of indoor air radon risk [27], which characterizes indoor air radon risk by county with 3 levels (Supporting Information Figure S3.1): Low (Zone 3), medium (Zone 2), and high (Zone 1) risk. Details on the calculation of the other confounding variables are available in the supporting information. Incidence risk ratio (IRR), or the ratio of the probabilities of disease when a given predictor variable is increased by one unit, is obtained for each variable by exponentiating its coefficient ($IRR = e^{\beta}$). We create and compare models with increasing levels of controlling for confounding variables. First, a crude model or model with only groundwater radon is produced. Second, in the *adjusted model* 1 we control for the effects of indoor air radon risk by including indoor air radon zones. Third, we

control for additional confounding factors including smoking, race, public water supply, and residential tenure in *adjusted model* 2. Lastly, we control for all confounders including gender and age with a *stratified model*.

Second, to utilize the point level exposure information from the groundwater Rn estimates[3] we conduct a logistic regression analysis on lung and stomach cancer cases that are assigned a 0/1 status based on their membership in a cluster [19,21]. This approach allows address level exposure information to be utilized in case-only studies and where a case-crossover study design is not sensible. Cancer clusters are identified by calculating the Anselin Local Moran's I on normalized excess case counts[21] $c_i = (o_i - e_i)/e_i$, where o_i is the number of observed cancer cases per census tract and $e_i$is the expected number of cases calculated as the North Carolina state average for the study period and gender and age adjusted for each census tract.These cancer clusters delineate geographic regions associated with unknown elevated risk factors. To identify these risk factors, we assign each individual cancer cases with a 0/1 binary variable *M* indicating their *membership* in cancer clusters. We model the probability that a lung or stomach cancer is a member of a cancer cluster using the logistic regression model

$$\text{logit}(M) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \varepsilon \tag{3.2}$$

where $logit(M)$ is the logit link function that transforms the binary membership dependent variable *M* to the appropriate scale for estimation, β_n are linear coefficients for the individual predictor variables X_n, and ε is the error term. We implement the logistic model using a GLM approach. Details of the logistic regression model are available in the supporting information. The variable X_n of interest represents the groundwater radon log-concentration log-pCi/L at the address of the cancer case, which we obtain via a spatial join from the estimated address-level groundwater radon estimates of Messier et al.[3]. The same confounding variables are included in the logistic model as in the NB2; however, differences due to the address-level information are present, which are explained in detail in the supporting information. The odds ratio (OR), or the ratio of the odds that a case is a member of a cluster when a given predictor variable is increased by one unit, is calculated for each variable by exponentiating the logistic regression model coefficient. Similarly to the NB2 model, we create and compare models with increasing levels of controlling for confounding variables; however, instead of stratification by gender and age, they are included as explanatory variables resulting in the *full model*.

Spatial auto-correlation of model residuals is assessed by examining a spatial covariance plot of the model Pearson residuals. If significant auto-correlation is present, which can potentially bias parameter and standard error estimates, then we implement a generalized estimating equation (GEE)[28–31], which accounts for correlations between clusters and assumes no correlation within clusters. GLM's are modeled using the COUNT package[25] and GEE's are modeled using the GEE package[32] of the R statistical software. Spatial covariance of residuals are calculated using the BMElib[33] numerical toolbox in MATLAB. The cluster analysis was performed using the *Cluster and Outlier Analysis* tool in ArcGIS 10.0[34].

Results

Lung Cancer

Results for the crude, adjusted 1, adjusted 2, and gender and age stratified lung cancer NB2 models are summarized in Table 3.2. The groundwater radon IRR for model adjusted 2 is positive and statistically significant (IRR =1.05, 95% CI=(1.01,1.08)). Residual spatial-autocorrelation in the lung cancer NB2 model is considered insignificant based on the Pearson covariance plots (Supporting Information Figure S3.2).

The state-wide observed incidence for lung cancer during the study period is 95.7 and 52.8 cases per 100,000 person-years for males and females respectively. This rate was used as the expected incidence in the cluster analysis of normalized excess cancer cases, which resulted in 254 out of 1554 (16.3%) census tracts with higher than expected rates of lung cancer (Supporting Information Figure S3.3). A total of 13,414 (21%) cases occur within the clusters.

Table 3.15. Lung Cancer Negative Binomial regression results for groundwater radon concentration for multiple models. The crude model contains only groundwater radon. Adjusted model 1 contains groundwater radon and is controlled for indoor air zones. Adjusted Model 2 contains model 1 plus all of the confounders except age and gender, which are smoking prevalence, residential tenure, percent public water, percent white race, and percent black race. The last model is stratified by gender and age. Results are expressed as IRR (95% Confidence Interval). ** Significant at 95% Confidence Interval. *Significant at 90% Confidence Interval. Groundwater radon unit = log-pCi/L averaged across census tracts. Indoor air radon zone is an ordinal variable with Rn Zone 3 as the reference level. Rn Zone 3 is the lowest risk of indoor air radon.

	Crude	**Adjusted 1**	**Adjusted 2**	**Males**		**Females**	
				Age <65	Age ≥ 65	Age <65	Age ≥ 65
Intercept	5.0e-4 (4.0e-4,7.0e-4)**	0.0006(0.0005, 0.0008)**	8.9e-5(6.0e-5,1.3e-4)**	3.4e-5 (2.1e-5,5.7e-5)**	0.0014(9.9e-4,0.002)**	3.5e-5 (2.1e-5,5.8e-5)**	0.0008(0.0005 ,0.001)**
Groundwater Radon	1.05 (1.02,1.08)**	1.02(0.99,1.06)	1.05(1.01, 1.08)**	1.01 (0.97, 1.06)	1.04 (1.02,1.07)**	1.06 (1.02,1.11)**	1.06 (1.03,1.10)**
Rn Zone 2		1.04(0.99,1.10)	0.94(0.89, 0.99)**	0.97 (0.90, 1.04)	0.95 (0.91,0.99)**	0.93 (0.87,1.00)*	0.95 (0.90,1.004)*
Rn Zone 1		1.19(1.09,1.31)**	1.06(0.97, 1.15)	1.01 (0.90, 1.13)	0.82 (0.76,0.88)**	0.96 (0.85,1.07)	0.89 (0.82,0.97)**

Table 3.3 summarizes the results of the crude, and confounder adjusted lung cancer logistic regression model for cluster membership. The fully adjusted address-level logistic GEE model indicates that groundwater radon exposure is a significant risk factor for the cluster membership of lung cancer cases (OR=1.32, 95% CI=(1.28,1.36)). Results for the confounding

variables in both models are available in the supporting information (Supporting Information Tables S3.1, S3.2). Residual spatial-auto-correlation in the lung cancer logistic model is considered insignificant based on the Pearson covariance plots (Supporting Information Figure S3.2).

Table 3.16. Lung cancer logistic GLM results representing the odds (OR, 95% Confidence Interval) a case is within the lung cancer cluster. The crude model contains only groundwater radon. Adjusted model 1 contains groundwater radon and is controlled for indoor air zones. Adjusted Model 2 contains model 1 plus all of the confounders except age and gender. The full model contains model 2 plus age and gender. ** Significant at 95% Confidence Interval. * Significant at 90% Confidence Interval. Groundwater radon unit = log-pCi/L. . Indoor air radon zone is an ordinal variable with Rn Zone 3 as the reference level. Rn Zone 3 is the lowest risk of indoor air radon.

	Crude	Adjusted 1	Adjusted 2	Full
Intercept	0.05 (0.04,0.06)**	0.05(0.04,0.06)**	0.005 (0.004,0.007)**	0.005(0.004,0.007)**
Groundwater Radon	1.30 (1.27,1.33)**	1.29(1.26,1.33)**	1.32 (1.28,1.36)**	1.32(1.28,1.36)**
Rn Zone 2		0.74(0.70,0.77)**	0.69 (0.65,0.72)**	0.69(0.65,0.72)**
Rn Zone 1		2.18(2.04,2.34)**	2.01 (1.87,2.16)**	2.00(1.87,2.15)**

Stomach Cancer

Groundwater radon IRR are generally positive, but insignificant in the crude, adjusted 1, adjusted 2 (IRR =1.02, 95% CI=(0.97,1.08)), and three out of four model stratifications each in the stomach cancer NB2 model. Full results are available in Supporting Information Table S3.3. The state-wide observed incidence for stomach cancer during the study period is 8.2 and 4.1 cases per 100,000 person-years for males and females respectively. This rate was used as the expected incidence in the cluster analysis of normalized excess cancer cases, which resulted in

113 out of 1554 (12.8%) census tracts with higher than expected rates of stomach cancer (Supporting Information Figure S3). A total of 667 (12.8%) cases occur within the clusters. Table 3.4 shows the GLM and GEE results for the stomach cancer crude and adjusted logistic model. The GEE is the best model because the GLM showed significant residual spatial auto-correlation; however, after implementing a GEE with a 3 by 3 exchangeable covariance structure[29], spatial auto-correlation was significantly reduced (Supporting Figure S3.2). The logistic GEE model indicates that groundwater radon exposure is a significant risk factor for cluster membership of stomach cancer cases (OR=1.18, 95% CI=1.07-1.31 for the full GEE model). Results for the confounding variables are in Supporting Information Tables S3.4.

Table 3.17. Stomach cancer logistic GLM and GEE results representing the odds (OR, 95% Confidence Interval) that a stomach cancer case falls within a local stomach cancer cluster. The crude model contains only groundwater radon. Adjusted model 1 contains groundwater radon and is controlled for indoor air zones. Adjusted Model 2 contains model 1 plus all of the confounders except age and gender. The full model contains model 2 plus age and gender. ** Significant at 95% Confidence Interval. * Significant at 90% Confidence Interval. Groundwater radon unit = log-pCi/L. . Indoor air radon zone is an ordinal variable with Rn Zone 3 as the reference level. Rn Zone 3 is the lowest risk of indoor air radon.

	GLM				GEE			
	Crude	**Adjusted 1**	**Adjusted 2**	**Full**	**Crude**	**Adjusted 1**	**Adjusted 2**	**Full**
Intercept	0.03 (0.01,0.05)**	0.03(0.01,0.06)**	0.016 (0.005, 0.05)**	0.016(0.005,0.05)**	0.009 (0.002, 0.04)**	0.02(0.005,0.09)**	0.011 (0.001, 0.09)**	0.011(0.001,0.10)**
Groundwater Radon	1.29 (1.18,1.42)**	1.29 (1.15,1.44)**	1.22 (1.08,1.36)**	1.21(1.08, 1.36)**	1.47 (1.24,1.75)**	1.22(1.09,1.38)**	1.19 (1.07,1.32)**	1.18(1.07, 1.31)**
Rn Zone 2		1.09 (0.89,1.33)	1.36 (1.09,1.68)**	1.36(1.09, 1.69)**		2.01(1.38,2.93)**	2.02 (1.23,3.31)**	2.03(1.23, 3.35)**

Rn Zone 1	0.80 (0.55,1.14)	1.16 (0.79,1.68)	1.15(0.78, 1.64)	3.56(1.88 ,6.7)**	3.12 (1.37,7.14)**	3.12(1.37, 7.1)**

Discussion

We presented ecological census tract and case-only individual level models for lung and stomach cancer in North Carolina, United States. Our goal was to quantify the associations between groundwater radon exposure and lung and stomach cancer, while not only considering the effects of known confounders, but also the spatial scale of outcome and explanatory variables. There has been several studies supporting that air radon is a significant risk for lung cancer [7,10–14] but there has been only one epidemiology study of groundwater radon exposure and lung cancer, and it was an ecological study for mortality[16] at the county level. There is general consensus on the biological and physical plausibility of groundwater radon leading to stomach cancer[8,9,15]; however, there has only been one epidemiology study with a small sample size and lack of confounders[4] to directly measure this association, which showed an insignificant association. Our study is the first epidemiological analysis finding a significant positive association between groundwater radon exposure and lung cancer *incidence* rates, and the first to find that an increase of 1 log-pCi/L in groundwater radon log-concentration significantly increases the odds that both lung cancer cases (OR=1.32, 95% CI=1.28-1.36) and stomach cancer cases (OR=1.18, 95% CI=1.07-1.31) are geographically clustered after controlling for confounding factors.

Groundwater radon is a source of indoor air radon due to radon's transfer from water to air during showers[42], laundry, and dishes[8]. We found groundwater radon concentrations to be a significant risk factor for lung cancer incidence rates consistently across all ecological NB2 models (Table 3.2). The crude model results in an IRR of 1.05 (95% CI, 1.02-1.08); moreover, we obtain the same IRR in the *adjusted model* 2. We further investigate risks by stratifying by age and gender, which results in groundwater radon as a significant risk factor in three out of four groups, with females at a slightly larger risk. Our NB2 model results for lung cancer provide the first epidemiological evidence of effect modification of gender on the association between groundwater radon exposure and lung cancer incidence rates, which is shown with an IRR of 1.06 for females in both age stratifications, and an IRR of 1.01 (95% CI, 0.97-1.06) and 1.04

(95% CI, 1.02-1.07) for males below 65 and 65 and above, respectively. Furthermore, this is consistent with lung cancer logistic GLM model, which finds that, everything else being the same, male lung cancer cases are at a reduced odds of being member of a lung cancer cluster compared to female lung cancer cases (Supporting Information Table S3.2) with an OR of 0.97. The effect of other confounding variables are generally consistent with the literature, and their interpretations are available in the supporting information.

We also find groundwater radon concentration to be a significant risk factor for the crude (OR=1.30) and adjusted (OR=1.32) logistic GLM models for lung cancer, thus for every 2.7 (natural or Euler's *e*) times increase in groundwater radon concentration after controlling for all confounding factors, there is thirty-two percent increase in the odds that a lung cancer case is member of a cluster. Since we have a case-only study design, the OR does not have the usual interpretation of an increase or decrease in odds of disease given an exposure; however, it does maintain an interpretation that reflects the underlying risk. In this two-stage analysis procedure, the statistically significant clusters delineate regions with underlying geographical risk factors for lung cancer, and the subsequent logistic regression analysis of case cluster membership indicates that increased groundwater concentration is one these risk factors since it has an OR significantly greater than one.. It follows that our logistic GLM result supplements our census-tract ecological study in providing the first epidemiological evidence that groundwater radon concentrations results in an increased risk of lung cancer; and more importantly, the logistic model shows this based on a fine grained model of exposure that captures the variability of address-level groundwater radon *within* each census-tracts, which is important for radon since it is known to have significant local variability. Overall, our results for groundwater radon and lung cancer associations provide epidemiological evidence and support the National Research Council[8] assessment of increased risk of lung cancer from groundwater radon exposure.

Lung cancer from indoor air radon exposure is the most well-studied target organ and pathway combination for radon[7,8,12–14,35–37]. There is a general consensus that residential exposure from indoor air radon increases risk of lung cancer. This result was not seen in our ecological NB2 model, which showed indoor air radon exposure having mixed controlling effects (Table 3.2). Conversely, our logistic model shows a significant protective effect for individuals in indoor air zone 2 versus zone 3 (OR=0.69), and a significant risk for individuals in indoor air zone 1 versus zone 3 (OR=2.0). Previous studies report a linear effect with no-

threshold(LNT) [7,8,37]; however, there is evidence that the LNT model is inconsistent with experimental data and biological plausability[38–40]. Possible explanations for intermediate air radon in zone 2 having a protective effect and high air radon in zone 3 being a risk for lung cancer include the possibility that the air radon/lung cancer dose-response is not LNT in combination with the fact that residents of indoor air zone 2 counties are more likely than zone 3 (low expected air radon) counties to obtain residential protective measures against vapor intrusion thus explaining the protective result of indoor air radon in zone 2 counties compared to zone 3 counties.

Results from the ecological NB2 models for stomach cancer are all mostly insignificant with five out of six IRR at least one or greater (Supporting Information Table S3.2). Contrarily, the address-level crude, adjusted, and full logistic models are significant. As previously mentioned, there is significant local variability in groundwater radon measurements that is likely diluted from areal averaging, and subsequently makes finding a significant effect in the ecological NB2 model more difficult. Additionally, the importance of accounting for residual spatial-autocorrelation is evidenced by the fact that there is a difference in groundwater radon OR between the adjusted logistic GLM and the adjusted logistic GEE for stomach cancer (Table 3.4).

The GEE model shows that groundwater radon exposure is a significant risk factor for stomach cancer with an 18% increased odds of stomach cancer membership in a cluster for every 2.8 times increase in concentration while controlling for all confounding factors. Our results provide the first epidemiological evidence that groundwater radon is a significant environmental risk factor underlying stomach cancer clusters, which supports the National Research Council[8] that groundwater radon is a significant risk for stomach cancer, but disputes Auviven et al. finding of no significant effects of radon exposure to stomach cancer[4]. Auviven et al. insignificant but protective findings for uranium also contradict the positive association Wilkinson et al.[41] found between uranium deposits and stomach cancer incidence. Furthermore, Kjelberg and Wiseman[17] found significant positive associations between indoor air radon and stomach cancer incidence.

In the full GEE model, we find that the controlling effect of air radon is consistent with a linear increase in the air radon/stomach cancer dose-response with an OR of 2.03(1.23,3.35) for zone 2 and an OR of 3.12(1.37,7.1) for zone 1 (Table 3.4). In contrast to lung cancer, where we

saw air radon having protective effect in zone 2, we see that air radon is risk for stomach cancer in zone 2. Indoor air radon, originating from groundwater and subsurface vapor intrusion, is trapped by protective mucous and cilia in the pharynx and tracheobronchial tree. It is often subsequently cleared via mucociliary action and then swallowed. This explains the large enough dose to the stomach to see effects, but also how natural protective mechanisms help create a low dose to the lungs. It is also important from a regulatory and remediation standpoint because methods for controlling indoor air radon such active and passive soil depressurization[8,43] may not work as effectively for eliminating the routes of exposure through groundwater.

Our study provides epidemiological evidence for the association between groundwater radon and lung cancer incidence. Additionally, our results support the association between groundwater radon exposure and stomach cancer, which has been understudied and has mixed results. Limitations of the NB2 models are normal for ecological studies, which includes assigning exposures to an analysis unit area when it is known the exposure varies significantly at the individual level. The logistic models improved upon this; however, there were still some controlling ecological level variables assigned to individual cancer cases, plus the addition of overall model parameters with the cluster analysis step decreases model parsimony. Nonetheless, our study should provide not only evidence of the associations, but the results needed to calculate the sufficient sample size needed to design a larger, individual-level epidemiological analysis such as a retrospective case-control or a prospective case-cohort study.

In summary, our study developed models for lung and stomach cancer associations with groundwater radon at the ecological scale with negative binomial regression and at the address-level with logistic regression of case membership in cancer clusters. We find the first epidemiological evidence of the association between groundwater radon exposure and increased risk of lung cancer incidence while controlling for confounders at the ecological-level and increased risk of lung cancer at an address-level. This is also the first epidemiological analysis to find groundwater radon to be a significant environmental risk factor underlying stomach cancer.

REFERENCES

(1) Loomis, D. P. Radon-222 Concentration and Aquifer Lithology in North Carolina. *Groundw. Monit. Remediat.* **1987**, 33–39.

(2) Yang, Q.; Smitherman, P. E.; Hess, C. T.; Culbertson, C. W.; Marvinney, R. G.; Smith, A. E.; Zheng, Y. Uranium and radon in private bedrock well water in Maine: geospatial analysis at two scales. *Environ. Sci. Technol.* **2014**, *48*, 4298–4306.

(3) Messier, K. P.; Campbell, T.; Bradley, P.; Serre, M. L. Estimation of Groundwater Radon in North Carolina using Land Use Regression and Bayesian Maximum Entropy. *Prep.* **2015**.

(4) Auvinen, A.; Salonen, L.; Pekkanen, J.; Pukkala, E.; Ilus, T.; Kurttio, P. Radon and other natural radionuclides in drinking water and risk of stomach cancer: a case-cohort study in Finland. *Int. J. Cancer* **2005**, *114*, 109–113.

(5) Zhu, H. C.; Charlet, J. M.; Poffijn, A. Radon risk mapping in southern Belgium: an application of geostatistical and GIS techniques. *Sci. Total Environ.* **2001**, *272*, 203–210.

(6) De Francesco, S.; Tommasone, F. P.; Cuoco, E.; Verrengia, G.; Tedesco, D. Radon hazard in shallow groundwaters: amplification and long term variability induced by rainfall. *Sci. Total Environ.* **2010**, *408*, 779–789.

(7) Darby, S.; Hill, D.; Auvinen, A.; Barros-Dios, J. M.; Baysson, H.; Bochicchio, F.; Deo, H.; Falk, R.; Forastiere, F.; Hakama, M.; et al. Radon in homes and risk of lung cancer: collaborative analysis of individual data from 13 European case-control studies. *BMJ* **2004**, *330*, 223.

(8) National Research Council. *Risk Assessment of Radon in Drinking Water*; Washington D.C., 1999.

(9) Campbell, T.; Mort, S.; Fong, F.; Crawford-Brown, D.; Vengosh, A.; Cornell, E.; Field, W. R. *North Carolina Radon-in-Water Advisory Committee Report*; Raleigh, North Carolina, 2011.

(10) Field, R. W.; Smith, B.; Steck, D.; Lynch, C. F. Residential radon exposure and lung cancer: variation in risk estimates using alternative exposure scenarios. *J. Expo. Anal. Environ. Epidemiol.* **2002**, *12*, 197–203.

(11) Kendall, G. M.; Smith, T. J. Doses to organs and tissues from radon and its decay products. *J. Radiol. Prot.* **2002**, *22*, 389–406.

(12) Lubin, J. H.; Boice, J. D. Lung cancer risk from residential radon: meta-analysis of eight epidemiologic studies. *J. Natl. Cancer Inst.* **1997**, *89*, 49–57.

(13) Field, R. W. Environmental Factors in Cancer: Radon. *Rev. Environ. Health* **2010**, *25*, 33–38.

(14) Krewski, D.; Lubin, J. H.; Zielinski, J. M.; Alavanja, M.; Catalan, V. S.; Field, R. W.; Klotz, J. B.; Letourneau, E. G.; Lynch, C. F.; Lyon, J. I.; et al. Residential Radon and Risk of Lung Cancer. *Epidemiology* **2005**, *16*, 137–145.

(15) Crawford-Brown, D. J. Cancer fatalities from waterborne radon (Rn-222). *Risk Anal.* **1991**, *11*, 135–143.

(16) Hess, C. T.; Weiffenbach, C. V.; Norton, S. A. Environmental Radon and Cancer Correlations in Maine. *Health Phys.* **1983**, *45*.

(17) Kjelberg, S.; Wiseman, J. S. The relationship of radon to gastrointestinal malignancies. *Am Surg* **1995**, *61*, 822–825.

(18) CDC. U.S. Cancer Statistics: An Interactive Atlas http://apps.nccd.cdc.gov/DCPC_INCA/DCPC_INCA.aspx.

(19) Gilbert, M.; Xiao, X.; Pfeiffer, D. U.; Epprecht, M.; Boles, S.; Czarnecki, C.; Chaitaweesub, P.; Kalpravidh, W.; Minh, P. Q.; Otte, M. J.; et al. Mapping H5N1 highly pathogenic avian influenza risk in Southeast Asia. *PNAS* **2008**, *105*, 4769–4774.

(20) Loth, L.; Gilbert, M.; Osmani, M. G.; Kalam, A. M.; Xiao, X. Risk factors and clusters of Highly Pathogenic Avian Influenza H5N1 outbreaks in Bangladesh. *Prev. Vet. Med.* **2010**, *96*, 104–113.

(21) Torbick, N.; Hession, S.; Stommel, E.; Caller, T. Mapping amyotrophic lateral sclerosis lake risk factors across northern New England. *Int. J. Health Geogr.* **2014**, *13*.

(22) Tsai, P.-J.; Lin, M.-L.; Chu, C.-M.; Perng, C.-H. Spatial autocorrelation analysis of health care hotspots in Taiwan in 2006. *BMC Public Health* **2009**, *9*.

(23) Allshouse, W. B.; Fitch, M. K.; Hampton, K. H.; Gesink, D. C.; Doherty, I. A.; Leone, P. A.; Serre, M. L.; Miller, W. C. Geomasking sensitive health data and privacy protection: an evaluation using an E911 database. *Geocarto Int.* **2010**, *25*, 443–452.

(24) Hampton, K. H.; Fitch, M. K.; Allshouse, W. B.; Doherty, I. a; Gesink, D. C.; Leone, P. a; Serre, M. L.; Miller, W. C. Mapping health data: improved privacy protection with donut method geomasking. *Am. J. Epidemiol.* **2010**, *172*, 1062–1069.

(25) Hilbe, J. M. *Negative Binomial Regression*; 2nd ed.; Cambridge University Press, 2011; pp. 2–14,134–136.

(26) Messier, K. P.; Jackson, L. E.; White, J. L.; Hilborn, E. D. Landscape risk factors for Lyme disease in the eastern broadleaf forest province of the Hudson River valley and the effect of explanatory data classification resolution. *Spat. Spatiotemporal. Epidemiol.* **2015**, *12*, 9–17.

(27) US Environmental Protection Agency. EPA Map of Radon Zones http://www.epa.gov/radon/zonemap.html.

(28) Liang, K.-Y.; Zeger, S. L. Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika* **1986**, *73*, 13–22.

(29) Carl, G.; Kühn, I. Analyzing spatial autocorrelation in species distributions using Gaussian and logit models. *Ecol. Modell.* **2007**, *207*, 159–170.

(30) Dormann, C. F.; McPherson, J. M.; Araújo, M. B.; Bivand, R.; Bolliger, J.; Carl, G.; Davies, R. G.; Hirzel, A.; Jetz, W.; Kissling, W. D.; et al. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography (Cop.).* **2007**, *30*, 609–628.

(31) Dormann, C. F. Assessing the validity of autologistic regression. *Ecol. Modell.* **2007**, *207*, 234–242.

(32) Carey, V. J.; Lumley, T.; Ripley, B. Generalized Estimation Equation Solver, 2012.

(33) Christakos, G.; Bogaert, P.; Serre, M. L. *Temporal GIS: Advanced Function for Field-Based Applications*; Springer: New York, NY, 2002.

(34) ESRI. ArcGIS, 2012.

(35) Vinson, D. S.; Campbell, T. R.; Vengosh, A. Radon transfer from groundwater used in showers to indoor air. *Appl. Geochemistry* **2008**, *23*, 2676–2685.

(36) Pavia, M.; Bianco, A.; Pileggi, C.; Angelillo, I. F. Meta-analysis of residential exposure to radon gas and lung cancer. *Bull. World Health Organ.* **2003**, *81*, 732–738.

(37) Darby, S. C.; Whitley, E.; Howe, G. R.; Sally, J.; Kusiak, R. A.; Lubin, J. H.; Howard, I.; Tirmarche, M.; Tomdsek, L.; Radford, P.; et al. Radon and Cancers Other Than Lung Cancer in Underground Miners : a Collaborative Analysis of 11 Studies. *J. Natl. Cancer Inst.* **1995**, *87*, 378–384.

(38) Field, R. W.; Withers, B. L. Occupational and environmental causes of lung cancer. *Clin. Chest Med.* **2012**, *33*, 681–703.

(39) Hooker, A. M.; Bhat, M.; Day, T. K.; Lane, J. M.; Swinburne, S. J.; Morley, A. A.; Sykes, P. J. The Linear No-Threshold Model does not Hold for Low-Dose Ionizing Radiation The Linear No-Threshold Model does not Hold for Low-Dose Ionizing Radiation. *Radiat. Res.* **2004**, *162*, 447–452.

(40) Tubiana, M.; Feinendegen, L. E.; Yang, C.; Kaminski, J. M. The Linear No-Threshold Relationship Is Inconsistent Experimental Data. *Radiology* **2009**, *251*, 13–22.

(41) Azzam, E. I.; de Toledo, S. M.; Raaphorst, G. P.; Mitchel, R. E. J. Low-Dose Ionizing Radiation Decreases the Frequency of Neoplastic Transformation to a Level below the Spontaneous Rate in C3H 10T1 / 2 Cells. *Radiat. Res.* **1996**, *146*, 369–373.

(42) Wilkinson, G. S. Gastric Cancer in New Mexico Counties with Significant Deposits of Uranium.pdf. *Arch. Environ. Health* **1985**.

(43) Henschel, D. B. henschel radon mitigation 1994.pdf. *Radiat. Prot. Dosimetry* **1994**.

Supporting Information for Chapter 3

Lung and Stomach Cancer Associations with Groundwater Radon in North Carolina, United States at Multiple Spatial Scales

Kyle P. Messier[†] *and Marc L. Serre*[†]*

Authors' Affiliation:

†Department of Environmental Science and Engineering, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC 27599, United States

***Corresponding Author:**

Marc L. Serre

Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, University of North Carolina, 1303 Michael Hooker Research Center, Chapel Hill, NC 27599

Phone: (919) 966-7014 Fax: (919) 966-7911

The Supporting Information includes 16 pages, 4 tables, 3 figures, and 9 references.

Confounding independent variables

The following independent confounding variables were used

- US EPA Indoor Air Radon Zones[1]: County level indoor air radon zone designations were assigned to the census tracts. Zone 1 are counties the highest risk potential with predicted average indoor radon greater than 4 pCi/L; Zone 2 are moderate risk counties with average indoor radon between 2 and 4 pCi/L; and Zone 3 are low potential counties with average indoor radon less than 2 pCi/L. Variables are coded as an ordinal variable with Zone 3 as the reference level.
- Smoking prevalence (% of population): Smoking is known to be associated with both lung and stomach cancer[2]. Since reliable smoking information for cases is not known, we utilize smoking prevalence estimates at the census tract level[3] to account for cancer risks associated with smoking
- Public water (% of population): Differences in water source based on public versus private supply are associated with diseases including acute gastrointestinal illnesses[4,5] and potentially cancers due to disinfection byproducts[6]. We use dasymetric mapping [7] to downscale county level estimates on population using public supplied water and domestic self-supplied water[8] to create a variable that is the percent of a census tract population using public water. This variable captures the confounding effect that the usage of public water has on cancer risks.
- Residential tenure (Years): The etiologically required time period to get cancer through a chronic environmental exposure is approximated at the census tract level with mean residential tenure. The American Community Survey, part of the US Census, obtains information on the average length a person has lived at that residence. For this study we calculated mean residential tenure as the difference between 2010 and the average year that residents moved into their current household, thus a larger value indicating less residential mobility. We use the residential tenure variable to capture the confounding effect that longer exposure has on cancer risks.
- Age and gender are controlled through model stratification. Age is stratified at 65. Race is controlled by including percent white and percent black variables from the census in the model.

The same census level estimates of smoking prevalence[3] and residential tenure used in the NB2 model are assigned to the cases for the logistic models. The cases' public water supply status were assigned via a spatial join with a comprehensive public water service area polygon[9], and modeled as a binary variable. Age and gender variables were created based on the case data and modeled as binary variables. Race was also based on the case data and modeled as a categorical variable.

Logistic Regression Model

We model the probability that a lung or stomach cancer will fall in a cancer cluster given a set of explanatory variables with a logistic regression model, or generalized linear model with binomial distribution assumption and logit function link. The basic form is as follows:

$$g(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \varepsilon$$

Where $g(Y)$ is the logit link function that transforms the binary dependent data to the appropriate scale for estimation, β_n are linear coefficients for the predictor variables, X_n and ε is the error term.

Model Coefficient Interpretations

Model coefficient interpretations are provided for the Lung stratified NB2, Lung full GLM, and Stomach full GEE. The full tables including the confounding variable coefficients are in this supporting information. For the stratified NB2 models, the Males 64 and under stratification is provided since the other three stratifications have similar interpretations to their respective counterpart.

Lung cancer full NB2:

1) Males 64 and under have a one percent increase in lung cancer risk for every 2.7 times increase in groundwater radon concentration, with all other confounding variables held constant; however, it is an insignificant increase in risk.

2) Males 64 and under in an indoor air radon zone 2 have a three percent decrease in lung cancer risk compared to those in indoor air radon zone 3, with all other confounding variables held constant; however, it is an insignificant decrease in risk.

3) Males 64 and under in an indoor air radon zone 1 have a 1 percent increase in lung cancer risk compared to those in indoor air radon zone 3, with all other confounding variables held constant; however, it is an insignificant increase in risk.

4) Males 64 and under have a greater than 100 times increase in lung cancer risk for every one percent increase in smoking prevalence within their census tract, with all other confounding variables held constant.
5) Males 64 and under have a 1 percent increase in lung cancer risk for every additional year of residential tenure, with all other confounding variables held constant.
6) Males 64 and under have a 3 percent decrease in lung cancer risk for every 10 percent increase in the population using public water supply, with all other confounding variables held constant.
7) Males 64 and under have a 23 percent increase in lung cancer risk for every 10 percent increase in Black race within their census tract, with all other confounding variables held constant.
8) Males 64 and under have a 12 percent increase in lung cancer risk for every 10 percent increase in White race within their census tract, with all other confounding variables held constant.

Lung Cancer full logistic GLM:

1) For every 2.7 times increase in groundwater radon concentration (pCi/L) there is a 32 percent increase in odds of having lung cancer within a lung cancer cluster, with all other confounding variables held constant.
2) People in indoor air radon zone 2 have a 31 percent decrease in the odds of lung cancer case membership within a lung cancer cluster compared to those in indoor air radon zone 3, with all other confounding variables held constant.
3) People in indoor air radon zone 1 have a 2 times increase in the odds of lung cancer case membership within a lung cancer cluster compared to those in indoor air radon zone 3, with all other confounding variables held constant.
4) People of white race have a 2.75 times increase in the odds of having lung cancer within a lung cancer cluster compared to people of non-black or white race, with all other confounding variables held constant.
5) People of black race have 2.36 times increase in the odds of having a lung cancer case within a lung cancer cluster compared to people of non-black or white race, with all other confounding variables held constant.

6) People on public water supply have a 63 percent increase in the odds of lung cancer case membership within a lung cancer cluster compared to people not on public water, with all other confounding variables held constant.
7) For every additional year of residential tenure, there is a three percent increase in the odds of lung cancer case membership within a lung cancer cluster, with all other confounding variables held constant.
8) For every one percent increase in a person's census tract smoking prevalence, there is a 20 times increase in the odds of lung cancer case membership within a lung cancer cluster, with all other confounding variables held constant.
9) Males have a three percent decrease in the odds of having lung cancer within a lung cancer cluster compared to women, when all other confounding variables are held constant.
10) People 65 and over have a six percent increase in the odds of having lung cancer within a lung cancer cluster compared to people under 65, when all other confounding variables are held constant.

Stomach cancer logistic GEE:

1) For every 2.7 times increase in groundwater radon concentration (pCi/L) there is an 18 percent increase in odds of stomach cancer case membership within a stomach cancer cluster, with all other confounding variables held constant.
2) People in indoor air radon zone 2 have a 2.03 times increased odds of stomach cancer case membership within a stomach cancer cluster compared to those in indoor air radon zone 3, with all other confounding variables held constant.
3) People in indoor air radon zone 1 have a 3.12 times increase odds of stomach cancer case membership within a lung cancer cluster compared to those in indoor air radon zone 3, with all other confounding variables held constant.
4) People of white race have a 2.26 times increase in odds to have a stomach cancer case with a stomach cancer cluster compared to people of non-black or white race, with all other confounding variables held constant; however, it is an insignificant increase in risk.
5) People of black race have a 3.6 times increase in odds to have a stomach cancer case with a stomach cancer cluster compared to people of non-black or white race, with all other confounding variables held constant.

6) People on public water supply have a 2.36 times increase odds of lung cancer case membership within a stomach cancer cluster compared to people not on public water, with all other confounding variables held constant.
7) For every additional year of residential tenure, there is a seven percent increase in odds of stomach cancer case membership within a stomach cancer cluster, with all other confounding variables held constant.
8) For every one percent increase in a person's census tract smoking prevalence, there is a 99 percent decrease in stomach cancer case membership within a stomach cancer cluster, with all other confounding variables held constant.
9) Males are thirteen percent less odds than females to have a stomach cancer case within a stomach cancer cluster, when all other confounding variables are held constant.
10) People 65 and over have a twenty-five percent increase in odds to have a stomach cancer case within a stomach cancer cluster, when all other confounding variables are held constant.

Figures

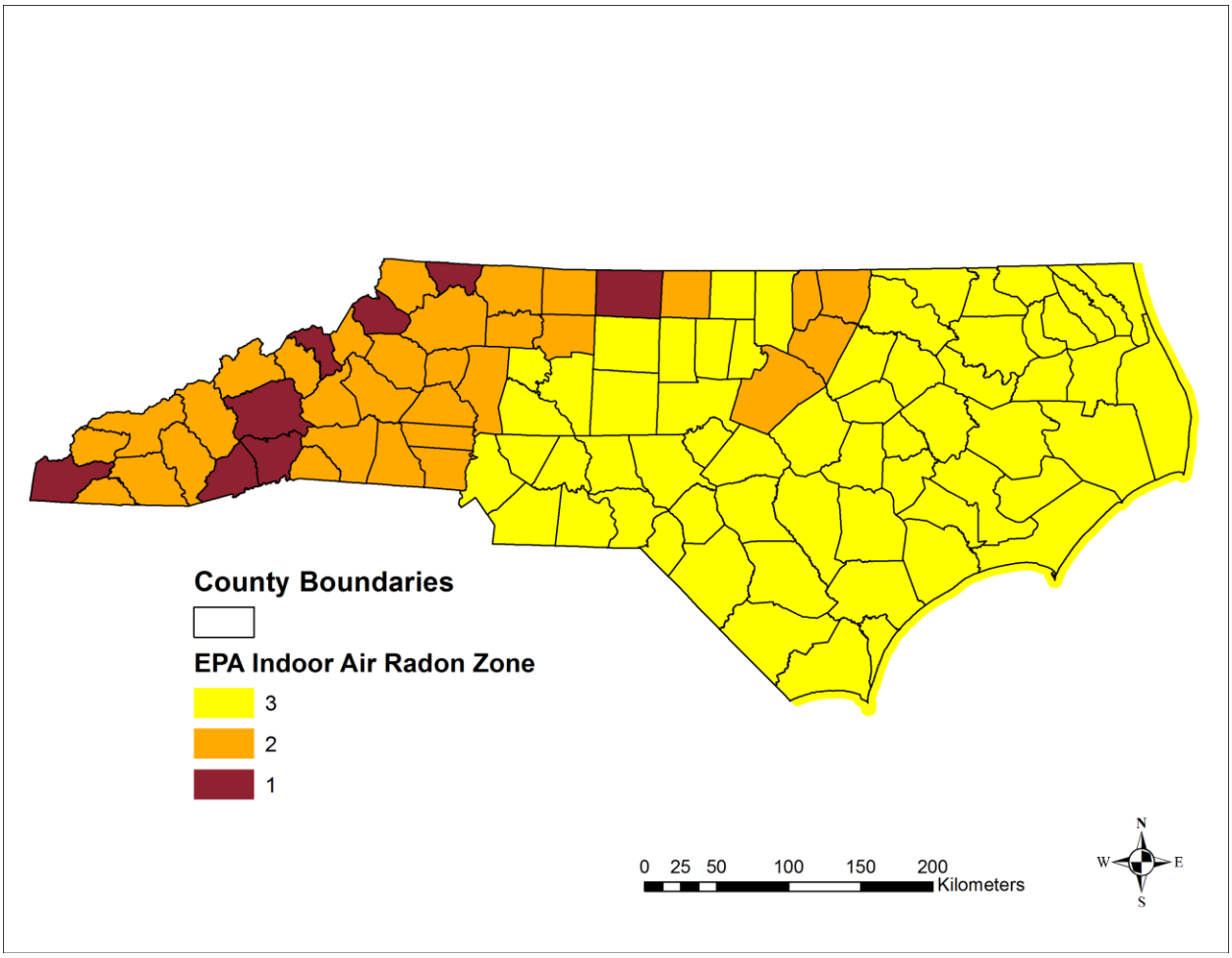


Figure S3.23. Indoor Air Radon risk zones by county as designated by the US Environmental Protection Agency[1]. Zone 1 (Highest Potential) counties have a predicted average indoor radon screening level greater than 4 pCi/L. Zone 2 (Moderate Potential) counties have a predicted average indoor radon screening level between 2 and 4 pCi/L. Zone 3 (Low Potential) counties have a predicted average indoor radon screening level less than 2 pCi/L.

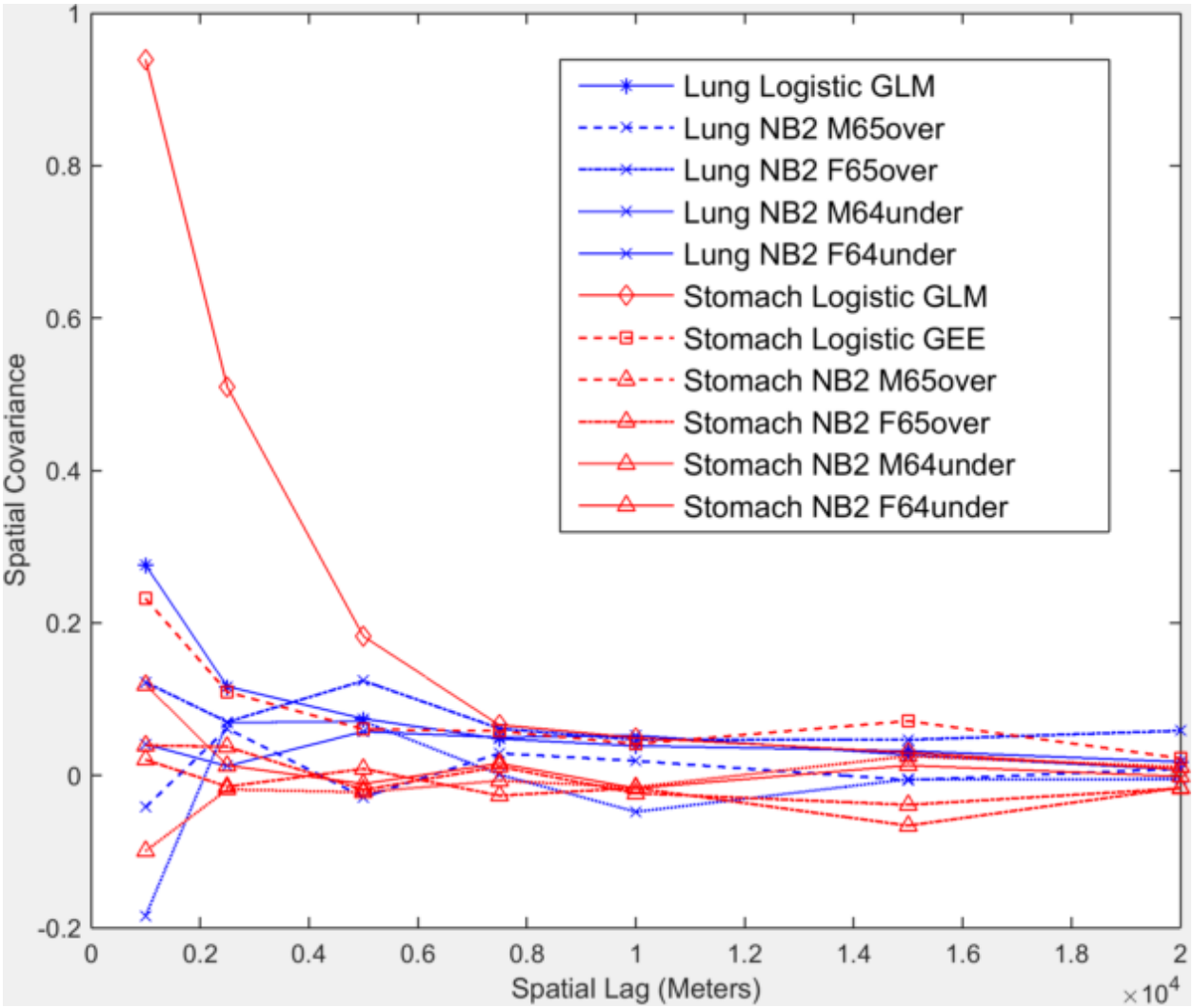


Figure S3.24. Pearson residual covariance plotted against spatial lags for all of the presented models. It is clear that the logistic GLM for stomach (red, diamond, solid line) cancer has significant spatial-autocorrelation in the residuals at short lags. A logistic GEE for stomach (red, square, dashed line) cancer is implemented which reduces residual spatial-autocorrelation to within the range of all other models.

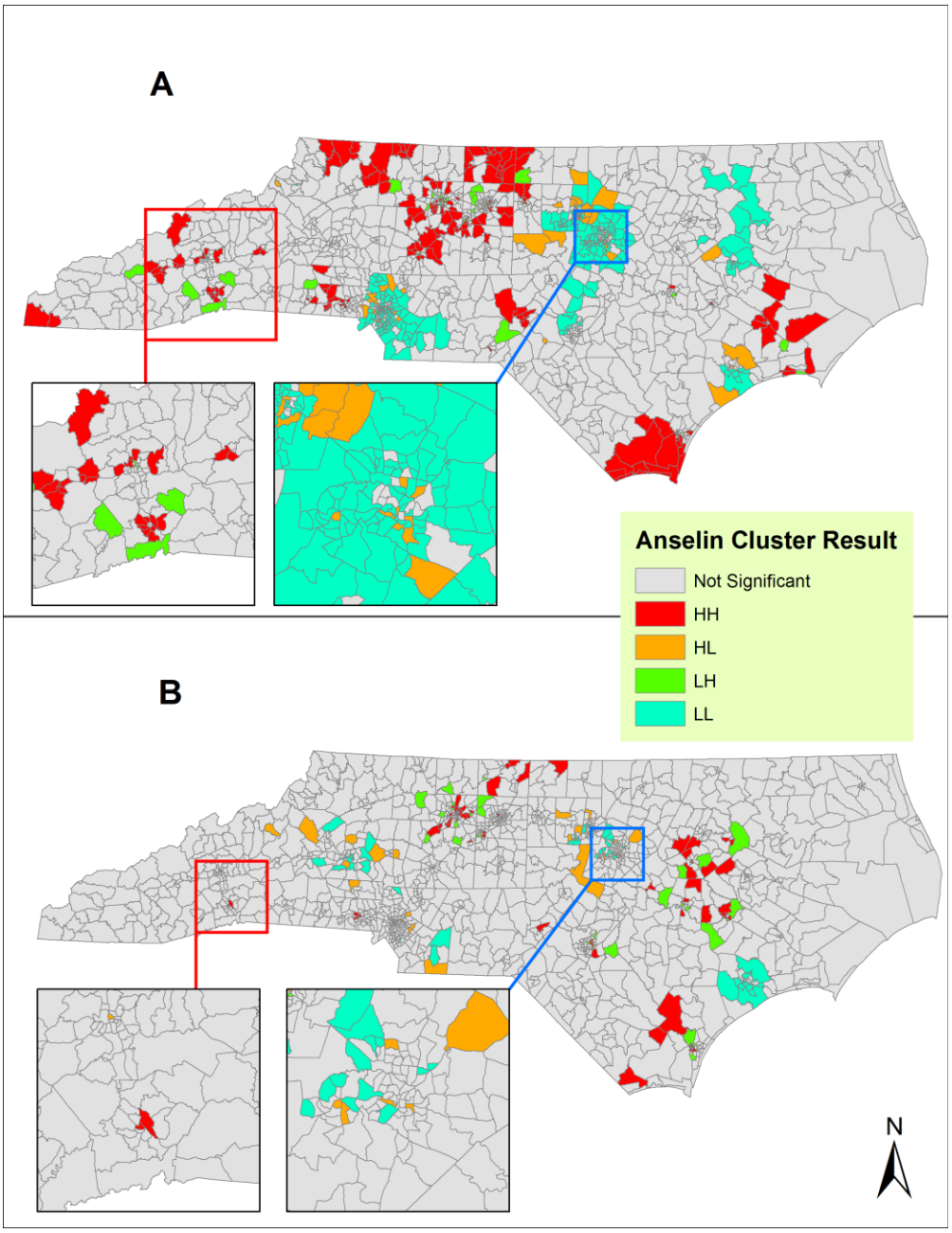


Figure S3.25. Anselin Local Moran's I clusters for excess, normalized A) Lung cancer, and B) Stomach cancer incidence calculated in ArcGIS 10.0. HH indicates statistically significant clusters of high values surrounding features of similar values. HL represents statistically significant clusters of high values surrounded by features with low values. LL represents statistically significant clusters of low values surrounded by low values. LH represents statistically significant low valued clusters next to other low values. Cases are assigned a 1 status if they within a census tract with value of HH or HL. All other cases are assigned a 0 status. Each

map has two inset maps of the Asheville (red border) and Raleigh (blue border) metropolitan areas.

Tables

Table S3.18. Lung cancer negative binomial regression models for groundwater radon and all confounding variables. The crude model contains only groundwater radon. Adjusted model 1 contains groundwater radon and is controlled for indoor air zones. Adjusted Model 2 contains model 1 plus all of the confounders except age and gender. The last model is stratified by gender and age and controls for all factors in adjusted model 2. ** = Significant at 95% Confidence Interval. * = Significant at 90% Confidence Interval. Units for groundwater radon are log-pCi/L. Units for the confounders are explained in previous text of the supporting information.

	Crude	**Adjusted 1**	**Adjusted 2**	**Males**		**Females**	
				Age <65	Age ≥ 65	Age <65	Age ≥ 65
Intercept	5.0e-4 (4.0e-4,7.0e-4)**	0.0006(0.0005,0.0008)**	8.9e-5(6.0e-5,1.3e-4)**	3.4e-5 (2.1e-5,5.7e-5)**	0.0014(9.9e-4,0.002)**	3.5e-5 (2.1e-5,5.8e-5)**	0.0008(0.0005,0.001)**
Groundwater Radon	1.05 (1.02,1.08)**	1.02(0.99,1.06)	1.05(1.01,1.08)**	1.01 (0.97,1.06)	1.04 (1.02,1.07)**	1.06 (1.02,1.11)**	1.06 (1.03,1.10)**
Rn Zone 2		1.04(0.99,1.10)	0.94(0.89,0.99)**	0.97 (0.90,1.04)	0.95 (0.91,0.99)**	0.93 (0.87,1.00)*	0.95 (0.90,1.004)*
Rn Zone 1		1.19(1.09,1.31)**	1.06(0.97,1.15)	1.01 (0.90,1.13)	0.82 (0.76,0.88)**	0.96 (0.85,1.07)	0.89 (0.82,0.97)**
Smoking			7.96 (4.97,12.7)**	105.0 (55.2,200)**	48.3 (32.2,72.7)**	51.6 (27.7,96.4)**	3.74 (2.31,6.06)**
Residen			1.01	1.02	0.98	0.994	0.96

tial Tenure	(1.005,1.02)**	(1.01,1.03)**	(0.97,0.99)**	(0.987, 1.01)	(0.956,0.97)**
Public Water Use (per 10%)	0.99 (0.97,0.999)**	0.97 (0.96,0.99)**	1.01 (0.99,1.02)*	0.99 (0.98,1.01)	1.03 (1.02,1.05)**
Black (per 10%)	1.17 (1.13,1.20)**	1.23 (1.19,1.28)**	1.05 (1.03,1.08)**	1.12 (1.08,1.17)**	1.03 (0.99,1.07)
White (per 10%)	1.15 (1.12,1.19)**	1.12 (1.08,1.16)**	1.03 (1.0004, 1.05)**	1.10 (1.06,1.15)**	1.08 (1.02,1.12)**

Table S3.19. Lung cancer logistic GLM results representing the odds (OR, 95% Confidence Interval) a case is within the lung cancer cluster for groundwater radon and all confounding variables. The crude model contains only groundwater radon. Adjusted model 1 contains groundwater radon and is controlled for indoor air zones. Adjusted Model 2 contains model 1 plus all of the confounders except age and gender. The full model contains model 2 plus age and gender. ** = Significant at 95% Confidence Interval. * = Significant at 90% Confidence Interval. Units for groundwater radon are log-pCi/L. Units for the confounders are explained in previous text of the supporting information.

	Crude	Adjusted 1	Adjusted 2	Full
Intercept	0.05 (0.04,0.06)**	0.05(0.04,0.06)**	0.005 (0.004,0.007)**	0.005(0.004,0.007)**
Groundwater Radon	1.30 (1.27,1.33)**	1.29(1.26,1.33)**	1.32 (1.28,1.36)**	1.32(1.28,1.36)**
Rn Zone 2		0.74(0.70,0.77)**	0.69 (0.65,0.72)**	0.69(0.65,0.72)**
Rn Zone 1		2.18(2.04,2.34)**	2.01 (1.87,2.16)**	2.00(1.87,2.15)**
White			2.76 (2.17,3.59)**	2.75(2.15,3.56)**
Black			2.35 (1.83,3.06)**	2.36(1.85,3.08)**
Public Water Supply			1.64 (1.56,1.72)**	1.63(1.56,1.71)**
Residential Tenure			1.03 (1.02,1.04)**	1.03(1.02,1.04)**
Smoking			19.8 (12.0,32.7)**	20.92(12.7,34.55)**
Male				0.97(0.93,1.01)*
65 Over				1.06(1.02,1.11)**

Table S3.20. Stomach cancer negative binomial regression models for groundwater radon and all confounding variables. The crude model contains only groundwater radon. Adjusted model 1 contains groundwater radon and is controlled for indoor air zones. Adjusted Model 2 contains model 1 plus all of the confounders except age and gender. The last model is stratified by gender and age and controls for all factors in adjusted model 2. ** = Significant at 95% Confidence Interval. * = Significant at 90% Confidence Interval. Units for groundwater radon are log-pCi/L. Units for the confounders are explained in previous text of the supporting information.

	Crude	Adjusted 1	Adjusted 2	Males		Females	
				Age <65	Age ≥ 65	Age <65	Age ≥ 65
Intercept	4.8e-5 (3.6e-5,6.5e-5)**	0.00005(3.2e-5,6.4e-5)**	1.5e-5(7.9e-6,3.0e-5)**	2.1e-5(6.1e-6,6.8e-5)**	0.0001(3.9e-5,3.2e-4)**	1.1e-5(2.8e-5,2.5e-6,4.6e-5)**	0.0001(3.6e-5,3.9e-4)**
Groundwater Radon	1.03 (0.99,1.08)	1.04(0.99, 1.10)	1.02(0.97,1.08)	0.95 (0.86,1.05)	1.09(1.01,1.18)**	1.00(0.87, 1.13)	1.01(0.91,1.11)
Rn Zone 2		0.95(0.88, 1.04)	1.0(0.92,1.09)	1.06 (0.90,1.24)	0.89(0.78,1.02)*	1.07(0.87, 1.32)	1.18(1.01,1.39)**
Rn Zone 1		1.04(0.90, 1.19)	1.12(0.97,1.29)	1.02 (0.76,1.34)	0.84(0.68,1.03)*	0.92(0.62, 1.34)	1.07(0.82,1.37)
Smoking			1.13(0.52,2.43)	1.35 (0.32,5.73)	3.45 (1.08,11.1)**	7.95 (1.25,50.98)**	4.34 (1.03,18.5)**
Residential Tenure			1.02(1.01,1.03)**	1.01 (0.99,1.03)	0.98 (0.97,1.00)*	1.01 (0.98,1.04)	0.97 (0.95,0.999)**
Public Water Use			0.99(0.97,1.01)	0.98 (0.95,1.02)	1.00 (0.97,1.02)	1.02 (0.98,1.07)	1.05 (1.01,1.09)**

(per 10%)					
Black (per 10%)	1.18(1.12,1.25)**	1.14 (1.04,1.26)**	1.12 (1.03,1.23)**	1.06 (0.96,1.18)	1.05 (0,96,1.16)
White (per 10%)	1.11(1.06,1.17)**	1.07 (0.98,1.17)	1.0 5(0.97,1.15)	0.98 (0.90,1.08)	0.98 (0.90,1.08)

Table S3.21. Stomach cancer logistic GLM and GEE results representing the odds (OR, 95% Confidence Interval) that a stomach cancer case falls within a local stomach cancer cluster for groundwater radon and all confounding variables. The crude model contains only groundwater radon. Adjusted model 1 contains groundwater radon and is controlled for indoor air zones. Adjusted Model 2 contains model 1 plus all of the confounders except age and gender. ** = Significant at 95% Confidence Interval. * = Significant at 90% Confidence Interval. Units for groundwater radon are log-pCi/L. Units for the confounders are explained in previous text of the supporting information.

	GLM				**GEE**			
	Crude	**Adjusted 1**	**Adjusted 2**	**Full**	**Crude**	**Adjusted 1**	**Adjusted 2**	**Full**
Intercept	0.03 (0.01,0.05)**	0.03(0.01,0.06)**	0.016 (0.005, 0.05)**	0.016(0.005,0.05)**	0.009 (0.002, 0.04)**	0.02(0.005,0.09)**	0.011 (0.001, 0.09)**	0.011(0.001,0.10)**
Groundwater Radon	1.29 (1.18,1.42)**	1.29 (1.15,1.44)**	1.22 (1.08,1.36)**	1.21(1.08, 1.36)**	1.47 (1.24,1.75)**	1.22(1.09,1.38)**	1.19 (1.07,1.32)**	1.18(1.07, 1.31)**
Rn Zone 2		1.09 (0.89,1.33)	1.36 (1.09,1.68)**	1.36(1.09, 1.69)**		2.01(1.38,2.93)**	2.02 (1.23,3.31)**	2.03(1.23, 3.35)**
Rn Zone 1		0.80 (0.55,1.1	1.16 (0.79,1.	1.15(0.78, 1.64)		3.56(1.88,6.7)**	3.12 (1.37,7.	3.12(1.37, 7.1)**

	4)	68)		14)**	
White	2.29 (1.22,4.9)**	2.19(1.17, 4.69)**		2.35 (0.84,6.6)	2.26(0.81, 6.30)
Black	3.69 (1.96,7.9)**	3.60(1.91, 7.73)**		3.65 (1.27,10.5)**	3.57(1.24, 10.3)**
Public Supply	2.28 (1.84,2.83)**	2.25(1.82, 2.81)**		2.39 (1.63,3.49)**	2.36(1.60, 3.47)**
Residential Tenure	1.05 (1.02,1.08)**	1.05(1.02, 1.08)**		1.07 (1.04,1.10)**	1.07(1.04, 1.10)**
Smoking	0.01 (0.001, 0.08)**	0.01(0.001,0.08)**		0.01 (0.001, 0.12)**	0.01(0.001,0.12)**
Male		0.89(0.75, 1.05)			0.87(0.80, 0.96)**
65 Over		1.26(1.05, 1.50)**			1.25(0.96, 1.62)*

REFERENCES

(1) US Environmental Protection Agency. EPA Map of Radon Zones http://www.epa.gov/radon/zonemap.html.

(2) *The Health Consequences of Smoking — 50 Years of Progress, A Report of the Surgeon General*; Atlanta, G.A., 2014.

(3) Ortega Hinojosa, A. M.; Davies, M. M.; Jarjour, S.; Burnett, R. T.; Mann, J. K.; Hughes, E.; Balmes, J. R.; Turner, M. C.; Jerrett, M. Developing small-area predictions for smoking and obesity prevalence in the United States for use in Environmental Public Health Tracking. *Environ. Res.* **2014**, *134*, 435–452.

(4) Blackburn, B. G.; Gunther, C. F.; Yoder, J. S.; Hill, V.; Calderon, R. L.; Chen, N.; Lee, S. H.; Levy, D. A.; Beach, M. J. *Surveillance for waterborne-disease outbreaks associated with drinking water - United States, 2001-2002*; 2004; Vol. 53, pp. 23–45.

(5) DeFelice, N.; Johnston, J. E.; Gibson, J. M. Burden of Acute Gastrointestinal Illness from Microbial Contaminants in North Carolina Community Water Systems. *Environ. Sci. Technol.*

(6) Villanueva, C. M.; Cantor, K. P.; Grimalt, J. O.; Malats, N.; Silverman, D.; Tardon, A.; Garcia-Closas, R.; Serra, C.; Carrato, A.; Castaño-Vinyals, G.; et al. Bladder cancer and exposure to water disinfection by-products through ingestion, bathing, showering, and swimming in pools. *Am. J. Epidemiol.* **2007**, *165*, 148–156.

(7) Sleeter, R.; Gould, M. *Geographic Information System Software to Remodel Population Data Using Dasymetric Mapping Methods*; Reston, Virginia, 2007.

(8) Kenny, J. F.; Barber, N. L.; Hutson, S. S.; Linsey, K. S.; Lovelace, J. K.; Maupin, M. A. Estimated Use of Water in the United States in 2005. *USGS Circ. 1344* **2005**.

(9) NCGICC. ncONEmap www.nconemap.com (accessed Jan 3, 2012).

APPENDIX: CONCLUSIONS, PUBLIC HEALTH RELEVANCE, AND FUTURE RESEARCH

Protecting public health is a paramount responsibility of environmental scientists and engineers. Through novel research scientists must develop and implement methods for risk assessments of contaminants harmful to human health. Land use regression (LUR) and Bayesian Maximum Entropy (BME) are both statistical modeling frameworks that can systematically and cost-effectively utilize publicly available datasets to in risk assessments. The work in these studies further developed these methods for exposure assessment and dose-response characterization of the deleterious human contaminants (NO_3^-) and radon (^{222}Rn).

Understanding the risk of groundwater NO_3^- and ^{222}Rn exposure is important because they are potential and known human carcinogens, respectively. The three studies in this work addressed the need for exposure assessment for groundwater NO_3^- and ^{222}Rn and the dose-response characterization for ^{222}Rn. The methodological developments and major findings in each study are detailed below and summarized in Table 4.1.

In chapter 1, *Nitrate Variability in Groundwater of North Carolina using Monitoring and Private Well Data Models,* we developed nonlinear LUR models for groundwater NO_3^- in shallow monitoring wells and deeper private wells. The nonlinear LUR models were novel because they were the first to quantify the spatial distribution of groundwater NO_3^- at a point-level spatial scale across a large domain. Literature-based and new explanatory variables were created that represented NO_3^- sources, attenuation, and transport factors. We developed a novel algorithm for selecting the best LUR model called *Constrained Forward Nonlinear Regression and Hyperparameter Optimization* (CFH-RHO) due to the nonlinear regression model in conjunction with the large amount of potential variables that were highly correlated. The final model selected by CFN-RHO showed that both wastewater treatment residual (WTR), or human waste biosolids sprayed on agricultural fields, and swine confined animal feeding operations (CAFOs) were both local sources of groundwater NO_3^- contamination, which had not yet been previously identified as sources in multivariable models. We then integrated the LUR model in the BME framework to produce the first space/time point-level estimates of groundwater NO_3^- including uncertainty estimates. A major finding from this result includes showing that groundwater NO_3^- in shallow monitoring wells in North Carolina is highly variable with many

areas predicted above the current human health standard of 10 mg/L. Contrarily, deeper private well model results show widespread, but low-level NO_3^- contamination. This finding is significant because of the human health implications, such as potential carcinogenic effects as low as 2.5 mg/L, but also for the ecological function as the deeper aquifer is potentially acting as a reserve of NO_3^- contamination to the surficial aquifer and surface waters. Another major finding from the novel point-level space/time mapping of groundwater NO_3^- was the elevated levels in the southeastern plains region of North Carolina due to the large amount of sources and the lack of subsurface attenuation factors.

In chapter 2, *Estimation of Groundwater Radon in North Carolina using Land Use Regression and Bayesian Maximum Entropy,* we developed an anisotropic geology-based LUR model for modeling spatial point-level groundwater ^{222}Rn, which was then integrated into the BME framework to produce the first point-level estimates including uncertainty characterization of groundwater ^{222}Rn across North Carolina. Geology and uranium based explanatory variables were created from the most up-to-date published geology data and from the United States Geological Survey (USGS) National Uranium Reconnaissance Survey (NURE). Variables were created to account for the anisotropic nature of geology. Major findings include mapping several areas across North Carolina's mountain and piedmont regions with elevated groundwater ^{222}Rn due to the underlying geology and uranium. Moreover, we performed non-parametric hypothesis tests on sediment Uranium concentrations within different areas of geological formations found to be associated with elevated ^{222}Rn and discovered that significant differences in the distributions of the sediment Uranium that are potentially showing intra-geological differences of observed ^{222}Rn.

Table 4.22. Summary of dissertation results.

Dissertation Chapter	Methodological Developments	Major Findings

1) Nitrate Variability in Groundwater of North Carolina using Monitoring and Private Well Data Models	• Nonlinear land use regression model for groundwater nitrate at the spatial point-level • Large variable space model selection algorithm for correlated variables in nonlinear regression	• Groundwater NO_3^- in monitoring wells that is highly variable with many areas predicted above the current standard of 10 mg/L • Groundwater NO_3^- in monitoring wells is elevated in the southeastern plains of North due to the larger amount of NO_3^- sources and the lack of subsurface attenuation factors • Both wastewater treatment residuals (WTR) and swine CAFOs were selected as local sources of groundwater NO_3^- contamination
2) Estimation of Groundwater Radon in North Carolina using Land Use Regression and Bayesian Maximum Entropy	• Accounting for geometric anisotropy through a land use regression model with ellipse based variables	• Several areas across the mountains and piedmont of North Carolina with elevated groundwater ^{222}Rn related to underlying geologic lithotectonic elements and Uranium • Sediment Uranium is potentially a diagnostic for intra-geological differences in groundwater ^{222}Rn.
3) Lung and Stomach Cancer Associations with Groundwater Radon in North Carolina, United States at Multiple Spatial Scales	• Case-only epidemiological analysis at the ecological and address-level scales	• Groundwater ^{222}Rn is a significant risk factor for lung cancer at the ecological and address-level. • Groundwater ^{222}Rn increases the odds of stomach cancer case membership in a stomach cancer cluster (OR=1.18) and lung cancer in lung cancer cluster (OR=1.32), after controlling for confounding factors.

In chapter 3, *Lung and Stomach Cancer Associations with Groundwater Radon in North Carolina, United States at Multiple Spatial Scales,* we utilize the exposure assessment of groundwater ^{222}Rn from chapter 2 to measure the dose-response of groundwater ^{222}Rn for the health outcomes of lung and stomach cancer. We had address geocoded and geomasked lung and stomach cancer cases for an eleven year period in North Carolina. Utilizing only case data, we developed ecological models, which examine the association between cancer incidence rates and areal averaged groundwater ^{222}Rn . Additionally, to utilize the point-level exposure estimates from chapter 2, we implemented a two-stage cluster analysis and then logistic regression of cases based on their membership within the cluster. This study was the first epidemiological analysis of the association between groundwater radon exposure and lung cancer, and the first to find a positive association between groundwater radon and stomach cancer. In the ecological models, we found groundwater ^{222}Rn to be a significant risk for lung cancer incidence in crude, confounder adjusted, and stratified models. In our address-level logistic regression model we found groundwater radon exposure to be a significant risk factor for stomach cancer (OR=1.18) and lung cancer (OR=1.32) after controlling for confounding factors.

Public Health Relevance

This body of research work provides advances in exposure assessment and dose-response methodology and practical real-world examples that can be used as resources for future protection of public health. The methods outlined in all three chapters utilize publicly available data that result in methodological developments and deliverable results in case-studies such as maps of predicted contaminant concentration. Environmental scientists, engineers, and regulatory agencies can all benefit from the methods demonstrated in this work while minimizing additional costs. Given that the methods are data-driven and that uncertainty is reduced in the neighborhood of monitoring locations, additional monitoring can result in more accurate exposure assessment implementing these methods; however, the uncertainty estimates provided in the exposure assessment can also help plan where additional monitoring efforts will have the largest marginal returns.

The LUR-BME framework implemented for the exposure assessments of groundwater NO_3^- and ^{222}Rn provide policy relevant information in two unique ways: First, the variables selected in the LUR provide information on environmental factors that are associated with the contaminants of concerns; and second, the maps of median concentration and error variance

provide evidence of the geographic distribution of high and low risk areas as they relate to the concentration and uncertainty of the contaminants.

The exposure assessment for groundwater NO_3^- has impacts for both human and ecological health. Excessive NO_3^- inputs into the environment can result in adverse changes to ecosystems such as eutrophication and harmful algal blooms. Groundwater and surface water systems are highly interconnected domains of our environment. In chapter 2, we recommended that the groundwater NO_3^- be utilized as the source (or sink) in a model for surface water NO_3^- such as the Soil and Water Assessment Tool (SWAT). The LUR-BME results of shallow aquifer, monitoring well NO_3^- are available for distribution to researchers as input into their models. In fact, within a month of publication, we were contacted by a consulting firm and *The Nature Conservancy* asking us to provide the results, which they are implementing in a SWAT model for a major basin in North Carolina.

The results of chapter 3, the epidemiological analysis of groundwater ^{222}Rn exposure and cancers of the lung and stomach, provide new and important information on the potential carcinogenic health effects of ^{222}Rn exposure. As previously mentioned, the general scientific consensus is that groundwater ^{222}Rn or drinking-water ^{222}Rn can cause stomach cancer; however, this was severely understudied as there was only one epidemiological analysis and it had insignificant results. Furthermore, the relationship between groundwater ^{222}Rn or drinking-water ^{222}Rn was also understudied as there was no study on that route of exposure and lung cancer incidence. Therefore, our results provide the only direct estimates of the dose-response relationship that can be used in future risk assessments. The results can also be used to provide accurate sample size calculations for a more expensive and time consuming individual level epidemiological analysis such as a case-control study. Lastly, the results are also important from a regulatory and remediation standpoint because methods for controlling indoor air radon such active and passive soil depressurization[8,43] may not work well for eliminating the routes of exposure through groundwater.

Future Research

The methodological developments and the major findings in this work provide a foundation for future exposure assessment and dose-response characterization; however, they also lead to more research questions and hypotheses. Examples of potential research questions based on the findings of these research studies are discussed below.

1) What variables if added to the groundwater NO_3^- models would increase its predictive ability? Results from chapter 1 show many significant explanatory variables in the LUR model; however, both models for monitoring and private wells have plenty of room for improvement. Records and timing of farm fertilizer applications including waste treatment residual applications could potentially resolve a significant amount of local scale spatial and temporal variability in groundwater NO_3^-. Additionally, information on private well pumping rates and decreasing the high detection limit of private well data could improve the private well model. All of these variables are possible; however, they would require primary data collection and a non-trivial upgrade in monitoring resources.

2) Can the LUR-BME model for groundwater NO_3^- be integrated into a multimedia model for NO_3^-? To further understand the impacts of legacy groundwater NO_3^- contamination on surface waters, a multimedia for NO_3^- could potentially be developed. The results could be included as a source in the Soil Water and Assessment Tool (SWAT) or the Spatially Referenced Regression on Watersheds (SPARROW) models. Additionally, to further understand the whole cycle, community multiscale air quality (CMAQ) model output for atmospheric nitrogen could also be considered in future research.

3) Does drinking-water radon cause an increase in the risk to develop lung or stomach cancer? While our study found significant and positive associations between groundwater ^{222}Rn and lung and stomach cancer, the study designs does not permit an interpretation about causality. To provide additional evidence towards a true causal relationship, a study would have to be designed with accurate spatial and temporal individual-level exposure information for a large cohort including cancers case and controls. A feasible study design would be a retrospective case-control study, whose sample size necessary to detect an effect can be calculated based on the results from our study. More detailed monitoring data would likely be necessary for both groundwater and indoor ^{222}Rn in order to accurately distinguish their exposures.

Lastly, the methods and results could be used in conjunction with population information to complete the process and produce the health-risk characterization. The benefit of our LUR-BME approach is that the estimates are characterized by a complete probability distribution function, which is essential to providing the overall uncertainty of a health-risk characterization. Potential outcomes of a health-risk characterization are estimates of lifetime probability of an

individual getting cancer based on their geographical location and the accompanied environmental exposures.